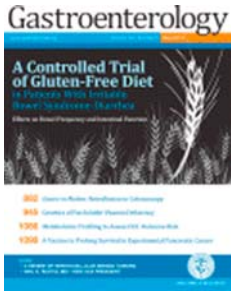
Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity

Wang, Jing; Mouradov, Dmitri; Wang, Xiaojing; Jorissen, Robert N.; Chambers, Matthew C.; Zimmerman, Lisa J.; Vasaikar, Suhas; Love, Christopher; Li, Shan; Lowes, Kym; Leuchowius, Karl-Johan; Jousset, Helene; Weinstock, Janet; Yau, Christopher; Mariadason, John; Shi, Zhiao; Ban, Yugan; Chen, Xi; Coffey, Robert J. C.; Slebos, Robbert J. C.

DOI:
10.1053/j.gastro.2017.06.008

Accepted Manuscript



Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity

Jing Wang, Dmitri Mouradov, Xiaojing Wang, Robert N. Jorissen, Matthew C. Chambers, Lisa J. Zimmerman, Suhas Vasaikar, Christopher G. Love, Shan Li, Kym Lowes, Karl-Johan Leuchowius, Helene Jousset, Janet Weinstock, Christopher Yau, John Mariadason, Zhiao Shi, Yuguan Ban, Xi Chen, Robert J.C. Coffey, Robbert J.C. Slebos, Antony W. Burgess, Daniel C. Liebler, Bing Zhang, Oliver M. Sieber

PII: S0016-5085(17)35742-6
DOI: 10.1053/j.gastro.2017.06.008
Reference: YGAST 61243

To appear in: *Gastroenterology*
Accepted Date: 12 June 2017

Please cite this article as: Wang J, Mouradov D, Wang X, Jorissen RN, Chambers MC, Zimmerman LJ, Vasaikar S, Love CG, Li S, Lowes K, Leuchowius K-J, Jousset H, Weinstock J, Yau C, Mariadason J, Shi Z, Ban Y, Chen X, Coffey RJC, Slebos RJC, Burgess AW, Liebler DC, Zhang B, Sieber OM, Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity, *Gastroenterology* (2017), doi: 10.1053/j.gastro.2017.06.008.

Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity

SHORT TITLE: Proteome-drug sensitivity associations in CRC

Jing Wang[1,2,*], Dmitri Mouradov[3,4,*], Xiaojing Wang[1,2,*], Robert N. Jorissen[3,4], Matthew C. Chambers[5], Lisa J. Zimmerman[5], Suhas Vasaikar[1,2], Christopher G. Love[3,4], Shan Li[3], Kym Lowes[3], Karl-Johan Leuchowius[3], Helene Jousset[3], Janet Weinstock[6], Christopher Yau[7,8], John Mariadason[9,10], Zhiao Shi[1], Yuguan Ban[11], Xi Chen[11,12], Robert J. C. Coffey[13,14], Robbert J.C. Slebos[15,], Antony W. Burgess[4,6,16], Daniel C. Liebler[5], Bing Zhang[1,2,**], Oliver M. Sieber[3,4,16,17,**]

[1]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA

[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

[3]Systems Biology and Personalised Medicine Division, The Walter and Eliza Hall Institute of Medial Research, Parkville, VIC 3052, Australia

[4]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3052, Australia

[5]Department of Biochemistry, Vanderbilt University, Nashville, TN 37232, USA

[6]Structural Biology Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia

[7]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom

[8]Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

[9]Olivia Newton-John Cancer Research Institute, Heidelberg, VIC 3084, Australia

[10]La Trobe University School of Cancer Medicine, Melbourne, VIC 3086, Australia

[11]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL 33136, USA

[12]Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL 33136, USA

[13]Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

[14]Veterans Affairs Medical Center, Nashville, TN 37212, USA

[15]Clinical Science Lab, Moffitt Cancer Center, Tampa, FL 33612, USA

[16]Department of Surgery, The University of Melbourne, Parkville, VIC 3052, Australia

[17]School of Biomedical Sciences, Monash University, Clayton, VIC 3800, Australia

**Joint first author

**Corresponding author

GRANT SUPPORT

This study was supported by the Ludwig Institute for Cancer Research, a NHMRC Project Grant (APP1079362), a Cancer Council Victoria Grant-in-Aid (APP1060964) and the Victorian Government's Operational Infrastructure Support Program, by National Cancer Institute (NCI) CPTAC awards U24CA159988 and U24CA210954, by NCI SPORE award P50CA095103, and by contract 15X038 from Leidos Biomedical Research, Inc. We also thank the support from NCI-Funded Special Programs of Research Excellence in GI Cancer. This research was supported by a Victorian Life Sciences Computation Initiative (VLSCI) grant number [VR0310, VR0311] on its Peak Computing Facility at the University of Melbourne, an

initiative of the Victorian Government, Australia. O.M.S. is a NHMRC R.D. Wright Biomedical Fellow (APP1062226). B.Z. is Cancer Prevention & Research Institutes of Texas (CPRIT RR160027) Scholar and McNair Medical Institute Scholar

CORRESPONDING AUTHORS:

Oliver Sieber, Systems Biology and Personalised Medicine Division, The Walter and Eliza Hall Institute of Medial Research, 1G Royal Parade, Parkville, VIC 3052, Australia. E-mail: sieber.o@wehi.edu.au

Bing Zhang, Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA. E-mail: bing.zhang@bcm.edu

CONFLICT OF INTEREST STATEMENT.

No conflict of interest or competing financial interests to disclose for all authors.

AUTHOR CONTRIBUTIONS

Conceptualization: AWB, BZ, DCL, OMS, RJCC, RJCS

Methodology: BZ, DCL, DM, RNJ, JW, OMS, XW

Software: CGL, CY, DM, JW, MCC, RJCS, RNJ, CGL, SV, XC, XW, YB, ZS

Validation: BZ, DM, JW, OMS, XW

Formal Analysis: BZ, DM, JW, OMS, RNJ, XW

Investigation: HJ, JW (Weinstock), KJL, KL, LJZ, SL

Resources: AWB, BZ, DCL, JM, OMS, ZS

Data curation: BZ, DM, JW, MCC, OMS, RJCS, XW

Writing – Original Drafts: BZ, DM, JW, OMS, XW

Write – Review & Editing: All authors

Visualization: BZ, DM, JW, OMS, XW

Supervision: BZ, OMS

Project Administration: AWB, BZ, DCL, OMS

Funding Acquisition: BZ, DCL, OMS, RJCC

ACCESSION NUMBERS: NCBI GEO - GSE90814, GSE90830

ACKNOWLEDGMENTS

We thank Doreen Agyapomaa for the preparation of compound screening plates and Eugene Kapp for providing access to the spike-in data set generated by the 2015 study of the Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (ABRF). We also thank the support from NCI-Funded Special Programs of Research Excellence in GI Cancer.

ABSTRACT

Background and Aims: Proteomics holds promise for individualizing cancer treatment. We analyzed to what extent the proteomic landscape of human colorectal cancer (CRC) is maintained in established CRC cell lines and the utility of proteomics for predicting therapeutic responses.

Methods: Proteomic and transcriptomic analyses were performed on 44 CRC cell lines, compared against primary CRCs (n=95) and normal tissues (n=60), and integrated with genomic and drug sensitivity data.

Results: Cell lines mirrored the proteomic aberrations of primary tumors, in particular for intrinsic programs. Tumor relationships of protein expression with DNA copy number aberrations and signatures of post-transcriptional regulation were recapitulated in cell lines. The five proteomic subtypes previously identified in tumors were represented among cell lines. Nonetheless, systematic differences between cell line and tumor proteomes were apparent, attributable to stroma, extrinsic signaling and growth conditions. Contribution of tumor stroma obscured signatures of DNA mismatch repair identified in cell lines with a hypermutation phenotype. Global proteomic data showed improved utility for predicting both known drug-target relationships and overall drug sensitivity as compared to genomic or transcriptomic measurements. Inhibition of targetable proteins associated with drug responses further identified corresponding synergistic or antagonistic drug combinations. Our data provide evidence for CRC proteomic subtype-specific drug responses.

Conclusions: Proteomes of established CRC cell line are representative of primary tumors. Proteomic data tend to exhibit improved prediction of drug sensitivity as compared to genomic

and transcriptomic profiles. Our integrative proteogenomic analysis highlights the potential of proteome profiling to inform personalized cancer medicine.

Keywords: colorectal cancer, cell lines, proteomics, drug sensitivity

INTRODUCTION

Studies of the genomic and transcriptomic landscapes of human colorectal cancer (CRC), have been instrumental in advancing our understanding of disease biology and the identification of clinically actionable aberrations [1-3]. While the major genomic and transcriptomic hallmarks and subtypes of CRC have been defined [4, 5], these explain only part of tumor clinical heterogeneity. The next challenge is to gain a detailed understanding of the dynamic protein pathways involved in normal and disease states, and we have recently characterized the proteome of primary CRCs from patients participating in The Cancer Genome Atlas (TCGA) project, identifying five major proteomic subtypes (Clinical Proteomic Tumor Analysis Consortium (CPTAC) [6]). From a therapeutic perspective, most drug targets are proteins rather than nucleic acids, and we have shown that DNA- or mRNA-level measurements are poor predictors of protein abundance [6].

Cancer cell lines are the most commonly utilized model systems in tumor biology and therapy development. Large cancer cell line-based projects, such as NCI-60 [7], Cancer Cell Line Encyclopedia (CCLE) [2] and Genomics of Drug Sensitivity in Cancer (GDSC) [3], have used molecularly heterogeneous cancer cell lines to identify stratification biomarkers and signatures for precision medicine. Nonetheless, controversy remains whether cell lines provide an appropriate representation of primary tumors, given the lack of organismal context, different growth conditions, and selection or acquisition of additional aberrations *in vitro*. Genomic analyses indicate that established cancer cell lines are suitable molecular proxies for primary tumors in many cancer types [2], yet findings at the transcriptomic level have been variable, with data for hepatocellular carcinoma [8] and colorectal cancer (CRC) [9] indicating similarity between cell lines and primary tumors, whilst data for breast cancer suggest pronounced differences [10].

Although some global proteomics data sets for cancer cell lines are available [11, 12] there exists no large-scale proteomic study comparing cell lines with primary tumors. It remains unknown whether cancer cell lines are representative of primary tumors at the proteome level and to what extent molecular programs and proteogenomic relationships are maintained *in vitro*. The relative utility of proteomic data as a predictor of anti-cancer drug responses in comparison to genomic and transcriptomic modalities has not been systematically investigated.

Here, we generated proteomic and transcriptomic data for a panel of 44 human CRC cell lines previously characterized at the genomic level [13]. We performed a comprehensive integrative proteogenomic analysis across these 44 cell lines and 95 CRCs and 60 normal tissue biopsies analyzed in our CPTAC project [6] to systematically evaluate cell lines as tumor models. We further integrated cell line proteogenomic data with drug sensitivity measurements to assess the utility of different types of omics data for predicting therapeutic responses and to connect tumor proteomic subtypes to drug sensitivity.

MATERIALS AND METHODS

CRC cell lines and primary tumors. A total of 44 CRC cell lines were studied (**Supplementary Table 1, Supplementary Methods**). In addition, we retrieved previously published genomic, transcriptomic and proteomic data on 95 primary tumor specimens from 90 CRC patients and proteomics data from 60 normal colon biopsies from 30 patients from our original CPTAC study [6], as well as RNA-Seq data for 48 normal colon and rectum samples deposited by the TCGA (**Supplementary Table 2-3**).

LC/MS-MS. The protein extraction and tryptic digestion of the frozen cell line pellets were performed as previously described for TCGA CRC specimens [6] (**Supplementary Methods**). Raw data for the cell lines, database search results, and the two versions of assemblies can be found at the Mass spectrometry Interactive Virtual Environment (MassIVE, ftp to massive.ucsd.edu, username: MSV000080374, password: a. HTTP access from the MassIVE website will be available after publication of the manuscript).

Transcriptome sequencing. RNA samples from CRC cell lines were extracted from pellets collected at the same time as the samples for proteomics analysis and sequenced to a depth of at least 28 million reads. Reads were subsequently aligned to human genome build Hg19 using Tophat (**Supplementary Methods**).

SNP microarray analysis. SNP array data on 38 cell lines from our cohort have been published previously [13]. SNP array assays on the additional DiFi, GEO, IS1, IS2, IS3 and V9P cells were performed at the Australian Genome Research Facility (AGRF) using CytoSNP-850K v1.1 and processed using OncoSNP v2.18 suite (**Supplementary Methods**).

Exome-capture sequencing. Whole exome mutation data on 35 CRC cell lines from our cohort have been published previously [13]. Additionally, DIFI, GEO, IS1, IS2, IS3, LIM1863, LIM2537, V9P and VAC05 cells were sequenced using the Nextera Rapid Capture Expanded Exome Enrichment Kit (Illumina) on an Illumina HiSeq 2000 System at the AGRF. Sequence alignment and calling of SNVs and INDEL in the absence of matched normal tissue were performed using

a hybrid of the GATK Germline and Somatic Best Practice Variant Detection Protocols (**Supplementary Methods**).

Variant peptide identification and analysis. To identify variant peptides, we derived customized protein sequence databases from matched WES and RNA-Seq data and then performed database searches using these databases for individual samples **(Supplementary Methods).**

VOOM/LIMMA analysis. Voom/limma analyses were performed using Limma and edgeR R packages, and method sensitivity and specificity for spectral count data were validated using the spike-in data set generated by the 2015 study of the Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (ABRF) (**Supplementary Fig. 1, Supplementary Methods**).

Online databases. The Human Protein Atlas, tumor stroma markers, KEGG pathways and GDSC (Genomics of Drug Sensitivity in Cancer) drug sensitivity data were downloaded from online resources **(Supplementary Methods).**

Correlation analysis. Spearman's correlation analysis of steady state mRNA and protein abundance, mRNA and protein variation, and relative mRNA-protein abundances required additional normalization steps that are outlined in **Supplementary Methods**.

Pathway signature identification. To assess whether genes in a given KEGG pathway had differing expression in tumors or cell lines relative to normal colorectal tissue, we modelled the protein or mRNA expression levels (cpm values for quantifiable genes) of pathway members using a linear mixed-effects model (lme4 R package) (**Supplementary Methods**).

Comparison of the impact of copy number alteration on protein abundance for cell lines and tumors. Evaluation of the association between copy number alteration and protein or mRNA levels were carried out using voom/limma analysis utilizing robust linear regression for gene-level log R ratios against protein or RNA-Seq expression levels **(Supplementary Methods).**

Comparison of the effect of candidate oncogene-targeting shRNAs on the proliferation of colon cancer cell lines. The shRNA gene level data was downloaded from the Achilles project website (https://portals.broadinstitute.org/achilles/datasets/5/download) and contained eight colon cancer cell lines overlapping with our 44 cell lines. We calculated the Spearman's correlation between shRNA score and log-transformed DNA copy number data across eight cell lines for each candidate oncogene (**Supplementary Methods**).

Drug sensitivity studies. Oxaliplatin (Cat# S1224), erlotinib (Cat# S7786) and regorafenib (Cat# S1178) were purchased from Selleck Chemicals. 5-fluorouracil (Cat# F6627) was obtained from Sigma. Cells were seeded into 384-well plates with compounds added to the cells in quadruplicate for 72hr. Cell viability was determined using CellTiter Glo 2 **(Supplementary**

Methods). For drug combination screening in HCT116 cells, 123 drugs were accessed from Compounds Australia, Griffith University, Australia (**Supplementary Tables 4-5**).

Comparison of omic modalities for prediction of drug sensitivity. Assessment of the utility of proteomic data for drug sensitivity prediction relative to mutation, DNA copy number, and mRNA expression data was undertaken using random forests and five-fold cross-validation for 5-fluorouracil, erlotinib, oxaliplatin, regorafenib and SN-38 over our panel of 44 CRC cell lines (**Supplementary Methods**).

Cell line proteomic and CMS subtype predictions. To assign cell lines to our previously identified proteomic subtypes [6], the R package *pamr* (http://CRAN.R-project.org/package=pamr) was used to apply our predefined signature genes from our CPTAC CRC tumor study to the cell line proteomic data (**Supplementary Methods**).

To assign CMS subtypes to cell lines and a dataset of 5 matched primary and metastatic tumor pairs (deposited in NCBI GEO: GSE90814), we used the CMSclassifier package in R (https://github.com/Sage-Bionetworks/CMSclassifier). (**Supplementary Methods**).

RESULTS

Proteomic analysis of CRC cell lines

We performed liquid chromatography-tandem mass spectrometry (LC-MS/MS) based shotgun proteomic analyses on 44 established CRC cell lines (**Supplementary Table 1, Supplementary Fig. 2)**, identifying a total of 10,643 distinct peptides (2,548,082 spectra) in an

assembly of 7,796 protein groups with a protein-level False Discovery Rate (FDR) of 4% (**Supplementary Table 6)**. To capture protein variants, we further searched customized protein sequence databases derived from matched whole exome sequencing (WES) and RNA-Seq data (**Supplementary Tables 7-8**). Out of 111,022 non-synonymous SNVs from RNA-Seq and WES data, we observed 1,702 unique variants at the proteomic level including 276 somatic variants reported in the TCGA/COSMIC databases and 952 germline variants listed in the Single Nucleotide Polymorphism Database (dbSNP) **(Supplementary data, Supplementary Table 9, Supplementary Fig. 3-4)**. The sparse detection of non-synonymous SNVs by peptide sequencing is consistent with our previous findings in primary tumors [6], reflecting the partial protein-coding sequence coverage achievable with the current proteomic technology.

Protein inventory concordance between cell line, tumor and normal samples

The cell line proteomic analysis was performed on the same platform previously used for the analysis of the TCGA tumors (n=95) and normal tissues (n=60) in our CPTAC project [6], and analysis of quality control samples across both projects demonstrated high platform stability (**Supplementary Fig. 5**). To determine the overlap between protein inventories of CRC cell line, tumor and normal colon samples, proteomic data were integrated into a joint assembly of 9,101 protein groups (**Supplementary Table 2)**. The protein inventory of cell lines was highly similar to those from tumor and normal tissues, exhibiting 98.0% and 90.9% overlap, respectively (**Supplementary Fig. 6a**). 103, 42 and 20 proteins were detected exclusively in cell line, tumor and normal samples, but most of these were low abundance proteins at the threshold of detection (**Supplementary Fig. 6b-6d**). Notably, proteome analysis detected 48% of the 18,178 protein-coding genes identified in corresponding mRNA datasets including the 44 matched cell line

samples, 87 matched tumor samples, and 48 normal samples (**Supplementary Table 3**), with similar representation of the major Gene Ontology (GO) categories (**Supplementary Fig. 7**). In the following analyses, we only used robustly quantifiable proteins, *i.e.*, proteins with a spectral count per million (CPM) >20 in ≥20% of samples.

Contribution of stroma components to tumor proteomes

Among the 4,904 quantifiable proteins from the CRC tumor and cell line proteomes, 747 (15.2%) exhibited significantly higher levels in the cell lines, whereas 979 (20.0%) displayed higher levels in the tumors (FDR<5% and fold-change >2, voom/limma, **Figure 1a**, **Supplementary Table 10**). Using GO enrichment analysis in WebGestalt [14], cell line-overexpressed proteins were enriched for cell growth and proliferation-related biological processes, such as metabolism and cell cycle, as anticipated for a comparison between *in vitro* cultured cell lines in log phase growth and primary tumor cells *in vivo* (**Figure 1b**, **Supplementary Table 11a and Supplementary Fig. 8a**). In contrast, tumor-overexpressed proteins were enriched for processes related to immune response, extracellular matrix, and response to extrinsic stimuli (**Figure 1b, Supplementary Table 11b,** and **Supplementary Fig. 7b**). The latter proteins also significantly overlapped with previously published cancer-associated fibroblast, leukocyte, or endothelial cell signatures [15] (p<2.2e-16, hypergeometric test, **Figure 1c**), indicating a substantial contribution of stroma to the tumor proteomes. Indeed, protein abundance for 82.3% of the tumor-overexpressed genes showed a negative correlation with tumor purity scores (ABSOLUTE algorithm [16]), in contrast to 38.4% among other genes (p<2.2e-16, two-sided Wilcoxon rank sum test, **Figure 1d** and **Supplementary Table 12**). We

also compared the mRNA profiles of cell lines and tumors and obtained similar results (**Supplementary Table 13-15** and **Supplementary Fig. 9-10**).

To characterize which components of the tumor stroma contributed to the tumor-overexpressed gene signature, we interrogated our tumor and cell line data for the expression of relevant stroma markers. Protein markers for blood plasma, extracellular matrix, endothelial cells, erythrocytes, fibroblasts, granulocytes, macrophages/monocytes, megakaryocytes/platelets and T cells were generally overexpressed (FDR<0.05 and fold-change>2, voom/limma) in the CRC samples as compared to the cell lines (**Figure 1e, Supplementary Table 16**). Analysis of RNA-Seq data additionally identified overexpression for markers of B lymphocytes and natural killer cells (**Supplementary Table 17**). Markers of the various tumor stroma components identified in the proteomics or RNA-Seq based analyses were verified by immunohistochemistry (IHC) data from the Human Protein Atlas (HPA)[17] (**Supplementary Fig. 11**). In contrast, IHC supported epithelial cell markers (EPCAM, KRT19, ITGA6, ITGB4; **Supplementary Fig. 12**), displayed similar expression levels (fold-change<2) in the cell lines and tumors (**Figure 1e, Supplementary Fig. 9e**).

To examine the impact of "contaminating" stroma on tumor proteome profiles, we intersected cell line and tumor data with the tumor-cell specific IHC expression scores from the HPA. Protein abundance measurements in tumor specimens showed only a weak concordance with corresponding IHC expression scores (p=0.075, Jonckheere's trend test), while cell line data exhibited a high level of concordance (p<2.2e-16) (**Figure 1f**).

Cell line proteomes reveal intrinsic biology of the hypermutation phenotype

To compare the utility of proteomic data from cell lines against that of tumor samples to elucidate cell-intrinsic molecular mechanisms, we investigated the protein profiles associated with the well-characterized hypermutation phenotype (both cohorts included 19 hypermutated cases; **Supplementary Table 1** and **Supplementary Fig. 2**). Using differential protein expression analysis followed by gene set enrichment analysis (GSEA) (**Supplementary Tables 18-20**), the DNA mismatch repair pathway was found to be significantly underexpressed in hypermutated cell lines compared with non-hypermutated cell lines (FDR=0.047), but this was not observed in tumors (**Figure 2a)**. Genes contributing to the statistical significance in the cell line data (blue bars, GSEA leading edge, **Figure 2b**) included the mismatch repair proteins MSH2 and MSH6, as well as two subunits of DNA polymerase delta, POLD1 and POLD2 (**Figure 2b** and **2c**). Loss of MSH2 and MSH6 expression are diagnostic of defective DNA mismatch repair, in particular for CRC associated with Lynch syndrome [18], and loss of POLD1 proof-reading function by somatic mutation in the exonuclease domain is implicated in causing tumor hypermutation phenotypes [19]. In contrast, tumor data associated the hypermutation phenotype with strong immune system signatures (**Figure 2a**), consistent with documented high levels of lymphocyte infiltration in hypermutated cases [20]. These results were replicated when examining mRNA-expression (**Supplementary Table 21-23, Supplementary Fig. 13**). Notably, MLH1 protein, loss of which underlies most hypermutated sporadic CRCs [18], was not detected in the proteomics data but was observed in the RNA-Seq data, with a greater dynamic signal range in cell lines relative to tumor samples (p=6.85e-05, Levene's test).

Tumor pathway signatures of post-transcriptional regulation are maintained in cell lines

We previously reported that mRNA and protein levels are only modestly correlated in the TCGA CRC cohort suggesting a major impact of post-transcriptional regulation [6], although omics analyses in tumor samples were performed on different specimen sections. To evaluate the relative contributions from biological and specimen variability, we compared mRNA-protein correlations in tumors samples with those from cell lines.

The average Spearman's correlation between steady-state mRNA and protein abundance within individual samples across genes was 0.60 for cell lines, compared to 0.46 for tumors (**Figure 3a**); the average correlation across samples within genes was 0.37 for cell lines, compared to 0.22 for tumors (**Figure 3b**). These results indicate that the tumor-based analyses likely have underestimated the protein-mRNA correlations, and emphasize the necessity of performing mRNA and protein measurements on the same tissue sample. Nevertheless, even for the cell line data, mRNA measurements remained poor predictors of protein abundance variations for many genes.

To investigate whether tumor signatures of post-transcriptional regulation at the biological pathway level were maintained in cell lines, we performed GSEA KEGG enrichment analysis on the average rank-differences between mRNA and protein expression (see Methods, **Figure 3c** and **Supplementary Table 24**). Tumors and cell lines exhibited high concordance for putative post-transcriptionally modulated pathways, with 66.7% of significant pathways overlapping between these cohorts (p<2.2e-16, hypergeometric test). Post-transcriptional up-regulation of protein expression was observed in both cohorts for 28 processes including 20

metabolic pathways, cAMP, cGMP signaling and cell adhesion functions. Only two pathways, p53 and Notch, showed evidence of coordinated post-transcriptional down-regulation.

Tumor intrinsic protein expression and pathway signatures are retained in CRC cell lines

To investigate to what extent proteome dysregulation in primary tumors was recapitulated in CRC cell lines, we compared protein abundances from cell line and tumor samples against those from normal samples. Cell lines and tumors exhibited a high correlation of expression changes relative to normal tissue (Spearman's correlation=0.66, p<2.2e-16; **Figure 4a, Supplementary Table 25**), with significant overlap between up-regulated and down-regulated proteins (FDR<0.05, fold change>2, voom/limma) identified for each group (p<2.2e-16, Fisher's exact test, **Figure 4b**). Nonetheless, expression in tumor samples tended to lie between that for normal tissues and cell lines, observed for 82.2% of the overlapping dysregulated proteins (p<2.2e-16, proportion test, **Figure 4c**), consistent with tumor samples representing an admixed population of neoplastic and normal cell types. Similar results were obtained when considering mRNA expression (**Supplementary Fig. 14a-14b, Supplementary Table 26**).

To gain a more detailed understanding of the level of conservation between cell lines and tumors at the level of protein pathways, we tested for coordinated protein expression changes within KEGG pathways as compared to normal tissue. Overall, changes in pathway expression were highly concordant between cell lines and tumors as observed at the individual protein level (Spearman's correlation=0.69, p<2.2e-16; **Supplementary Fig. 15a, Supplementary Table 27**). In particular, significant "intrinsic" pathways (FDR<0.05 for either group comparisons; left

panel of **Figure 4d**) related to genetic information processing and metabolism showed a high consistency of protein expression between tumors (purple points) and cell lines (orange points) compared to normal tissues (green points), with tumor pathway expression levels generally between cell lines and normal tissues. However, for "extrinsic" and stroma-related pathways including environmental information processing, cellular and immune-system related processes, tumors were more similar to normal tissues, while expression in cell lines was markedly decreased. These global protein and pathway category patterns were recapitulated for RNA-Seq data (**Supplementary Fig. 14c, Supplementary Fig. 15b, Supplementary Table 28**).

Influence of copy number aberrations on protein abundance across cell lines and tumors

While the impact of DNA copy number on mRNA expression is well established [4], our previous analysis of TCGA tumors suggested that this impact is less apparent with respect to protein expression [6]. To compare the effect of DNA copy number aberrations on gene expression between tumors and CRC cell lines, we retrieved DNA copy number states from matched SNP array data. DNA copy-number spectra in cell lines closely resembled those seen in tumors, with the most commonly gained chromosome arms including chromosome 7, 8q, 13, and 20q, and the most common deleted regions including 8p, 17p, and 18q (% gain = red bars, % loss = blue bars in **Figure 5a**). Overall, 989 proteins in CRC cell lines and 1524 proteins in tumors were associated with DNA copy-number changes (FDR<0.2, voom/limma, see Methods), with strengths of associations tracking with the respective frequencies of DNA copy number loss or gain (**Figure 5b, Supplementary Table 29-30**). As expected, similar but more pronounced results were found when analyzing associations between DNA copy number aberrations and mRNA expression (**Supplementary Table 31-32**). 438 protein-DNA measurement relationships

were detected across both tumors and cell lines (p<2.2e-16 for overlap, hypergeometric test; large points in **Figure 5b**), 90.0% of which also were detected at the mRNA level (green/red points). Among these proteins, 26 are known or proposed cancer genes (red points in **Figure 5b**, **Supplementary Table 33**). Proteins identified in regions of gain included the established EGFR oncogene [21], and candidates such as FOXK1, a forkhead transcription factor, and CNDP2, an activator of MAPK pathways. Increased expression of FOXK1 has been shown to promote CRC invasion and metastasis [22], and up-regulation of CNDP2 to facilitate colon cancer proliferation [23]. In regions of loss, we identified several putative tumor suppressors, including MTHFD1, a 1-tetrahydrofolate synthase. MTHFD1 deficiency has been shown to increase intestinal tumor incidence, number and burden in transgenic mouse models [24]. Using shRNA knockdown data from the Achilles project [25], we further validated six oncogene candidates (USP39, PARP1, EGFR, DLD, SRI and IDH3B) as "essential" genes in CRC (**Supplementary Fig. 16**).

Proteomics data better predicts CRC drug sensitivity

To evaluate the relative utility of proteomics data as a marker of drug responses in comparison to mutation, DNA copy number, and mRNA expression data, we retrieved response profiles for 210 drugs from the GDSC database which comprised 18 cell lines from our CRC cell line panel.

Considering 191 known drug-target gene associations quantifiable at the protein level (**Supplementary Table 34**), proteomics data identified 16.2% of the relationships (FDR <0.2), as compared to only 6.3% for mRNA, 5.3% for DNA copy number and 1.9% for mutation data (**Supplementary Table 35**). Among the significant drug-target gene pairs detected at the protein

level were multiple associations for EGF receptor family members (afatinib, cetuximab, gefitinib), heat shock protein 90 (CCT018159, SNX-2112) and β-tubulin family members (docetaxel, epothilone B, vinblastine, vinorelbine) (**Figure 6a**). Among pairwise comparisons in which at least one omics modality showed a discernable association (FDR <0.2), proteomic data showed greater correlations with drug responses than mRNA and DNA copy number data for 77.1% (27/35, *p*=0.001) and 81.6% (31/38, *p*=9.53e-05) of cases, respectively. Mutation data could not formally be evaluated for this latter comparison as only two quantifiable cases were significant in the overlap with the proteomics data.

Extending our association analyses to known drug-pathway relationships, proteomics data again identified more relationships (52.8%) than mRNA (25.2%), DNA copy number (1.6%) and mutation data (0%) (**Figure 6b, Supplementary Table 36**). The KEGG DNA replication (e.g. mitomycin C, SN-38), MAPK (e.g. TAK-715, trametinib) and PI3K-Akt (GDC0941, KIN001-102) pathways were among the significant drug-pathway pairs detected at the protein level (**Figure 6b)**. Similarly, for pairwise comparisons, proteomic data showed greater correlations with drug response than mRNA, DNA copy number and mutation data for 74.3% (55/74, *p*=2.36e-05), 97.0% (65/67, *p*=1.80e-14) and 100% (62/62, *p*=1.65e-13) of respective cases. In addition to the established drug-target relationships, responses for many drugs were correlated with protein signatures reflective of cell doubling rate (**Supplementary Fig. 17-18, Supplementary Tables 37-38**).

To more formally assess the utility of proteomic data for prediction of drug sensitivity relative to mRNA expression, DNA copy number and mutation data, we evaluated predictive

models using random forests and five-fold cross-validation. Given the limited number of CRC cell lines with matched GDSC data, we screened our 44 CRC cell lines panel against four major drugs used in the treatment of human CRC, 5-fluoruracil (5-FU), oxaliplatin, SN-38 and regorafenib. In addition, we tested the small molecule inhibitor erlotinib as a proxy for anti-EGFR antibody therapeutics (**Supplementary Table 39**). Significant correlations were observed between GDSC and our drug sensitivity data for two overlapping drugs, 5-FU and SN-38 (**Supplementary Fig. 19**).

As shown in Fig 6c, proteomics data demonstrated better performance for predicting sensitivity to 5-FU, SN-38, erlotinib, regorafenib and oxaliplatin in 11 out of 15 pair-wise comparisons against other modalities. Notably, proteomics data displayed a striking advantage for 5-FU, SN-38 and erlotinib. For regorafenib and oxaliplatin, only mutation data (yellow) and mRNA data (blue) outperformed proteomics data, respectively. In general, proteomics data thus provides an improved ability to predict the drug sensitivity of the CRC cell lines.

Proteins associated with drug sensitivity may be functionally implicated in determining drug responses. Pharmacological inhibition of targetable proteins contributing to drug resistance may synergize with baseline treatment, whereas inhibition of proteins conferring sensitivity may be antagonistic. To test this contention, we assembled 60 and 92 drugs whose inhibitory profiles included targetable protein implicated in responses to 5-FU or SN-38 (the active metabolite of irinotecan) (FDR <0.2 in GDSC, 48 and 56 targets), respectively, two mainstay treatments for CRC (**Supplementary Tables 4-5**). Dose-response curves for the inhibitor panel were determined for HCT116 colon cancer cells in the presence or absence of 5-FU or SN-38 at

IC30/40 concentrations, and drug combinations evaluated for evidence of synergy or antagonism based on excess over the Bliss (EOB) independence model. For both 5-FU and SN-38 treatment, EOBs tended to differ between drugs targeting protein markers of resistance as compared to markers of sensitivity (5-FU, p=0.011, SN-38, p=0.103, t-test), with the expected propensities to synergy or antagonism (**Figure 6d-e, Supplementary Tables 4-5**). For 5-FU treatment, inhibition with disulfiram, an efficacious ALDH inhibitor (incl. ALDH1 and ALDH2), was the top synergistic combination detected (**Figure 6f**). ALDH is a family of enzymes that play a key role in the metabolism of aldehydes and have been shown to oxidize and inactivate several prominent chemotherapeutic drugs [26]. ALDH activity has been associated with colon cancer resistance to irradiation and 5-FU [27]. Accordingly, disulfiram has previously been shown to potentiate gemcitabine and 5-FU treatment in colon cancer cells [28, 29].

Danusertib, an inhibitor against for Aurora A/B/C was identified as another synergistic compound with 5-FU (**Figure 6f**), and multiple inhibitors of Aurora kinases have been evaluated for the treatment of CRC in combination with 5-FU, with several in clinical trials [30]. Consistent with our findings, several studies have indicated that overexpression of Aurora kinases has a role in chemo- and radiotherapy resistance of CRC [31, 32].

For SN-38, combination with multiple tubulin inhibitors showed evidence of antagonism (**Figure 6f**). It has previously been reported that a primary mechanism of tubulin inhibitor resistance is simultaneous administration of a compound that inhibits cell cycle progression at the G2-M phase, the main phase of action of SN-38 [33].

Cell lines connect proteomic subtypes to drug sensitivity

Colorectal tumors can be classified into five proteomic subtypes that are largely distinct from the established transcriptomic subtypes [6]. Using a PAM prediction model trained on the primary tumor samples (**Supplementary Fig. 20**), 40 cell lines were assigned to a proteomic subtype with a prediction probability of >0.8 (**Supplementary Fig. 21)**, with representative cell lines identified for all five subtype classes (A-E) (**Figure 7a, Supplementary Table 1**). Cell lines were further categorized into transcriptomic subtypes using the CMSclassifier algorithm [5]. Subtypes CMS1, CMS2 and CMS3 were identified among cell lines, but subtype CMS4 was not assigned (**Supplementary Fig. 22a, Supplementary Table 1**). The failure to detect CMS4-assigned cell lines may be coincidental given our limited cohort size, or perhaps reflect the observation that this subtype signature is largely dominated by signals from tumor stroma [15, 34].

Comparing cell lines and tumors, proteomic and CMS subtypes were associated with similar distributions of genomic hallmarks across the cohorts, including MSI and CIMP status and mutations in *BRAF*, *APC*, *TP53* and *KRAS* (**Figure 7b, Supplementary Fig. 22b**). Interestingly, analysis of paired cell lines derived from the same tumor or primary tumor and metastatic derivatives identified some discordant assignments for proteomic subtypes. Discordances were also observed for transcriptomic subtypes, suggesting that proteomic and transcriptomic subtypes may represent transient states, with tumors adopting different subtypes with clonal evolution (**Supplementary Data**). Consistent with this suggestion, mutational differences were evident between paired cell lines at the genetic level (**Supplementary Data**). The transient nature of expression-based subtypes was further supported by microarray analysis for 5 matched primary tumors and liver metastases identifying discordant CMS subtypes for 3 of these pairs (**Supplementary Table 40)**.

To evaluate the potential value of tumor proteomic subtypes to predict drug response, we analysed the 5-FU, oxaliplatin, SN-38, regorafenib and erlotinib data for our 44 CRC cell line panel. GDSC data were not evaluated, due to the small cohort size. Although the number of cell lines in each subgroup were limited, proteomics subtypes were significantly associated with response to 5-FU, with subtype C exhibiting the greatest sensitivity (univariate p=0.014, ANOVA, **Figure 7c**). The association of proteomic subtypes with 5-FU response remained significant when adjusting for cell doubling time (adjusted p=0.003 ANOVA), which itself was directly related to 5-FU sensitivity (adjusted p=0.0007, ANOVA), or when the analysis was limited to microsatellite stable cases (univariate p=0.031, adjusted p=0.009, ANOVA, **Supplementary Fig. 23**). In contrast, transcriptomic subtypes showed no significant associations with drug response (**Supplementary Fig. 22c)**. The mechanism underlying the increased sensitivity of proteomics subtype C to 5-FU remains to be elucidated, but may be related to differences in 5-FU metabolic activation, detoxification or drug export [35, 36]. These results further underscore the potential of proteomic data for drug response prediction, in line with our protein signature and pathway analyses.

DISCUSSION

Our global proteomic characterization firstly demonstrates that CRC cell line proteomes maintain the major cell-intrinsic molecular programs, proteogenomic relationships and proteomic subtypes observed in primary tumors, highlighting the utility of cell lines as models for tumor biology, biomarker discovery and therapeutics. Most proteome aberrations and intrinsic pathway

signatures (*e.g.* genetic information processing and metabolism) showed concordant differences in both cell lines and tumors as compared to normal tissues. Relationships between protein expression and somatic DNA copy number changes in primary tumors were recapitulated in cell lines, identifying both established (EGFR) and candidate cancer genes (*e.g.* FOXK1, CNDP2 and MTHFD1) [22-24]. Integration of proteomic and transcriptomic data indicates that tumor post-transcriptional regulation at the biological pathway level is maintained in cell lines. The five proteomic subtypes previously identified for primary tumors were represented among cell lines and showed similar distributions of established genomic hallmarks. Notably, some heterogeneity in proteomic subtype assignments was observed between paired cell lines, as for transcriptomic subtypes, suggesting that expression-based subtype signatures may represent transient states.

Nonetheless, systematic differences between cell line and tumor proteomes were apparent, with major changes attributable to tumor stroma, extrinsic signaling and different growth conditions. Because of the significant contribution of the tumor stroma, the anticipated signatures of DNA mismatch repair and DNA proofreading polymerases identified in cell lines with a hypermutation phenotype were not detectable in the primary tumors. Instead, the proteomes of hypermutated tumors were characterized by signatures of immune infiltrates that are typically associated with such cases [20]. Overall, protein abundance measurements in cell lines showed a higher concordance with tumor-cell specific IHC expression measurements than did proteome profiles of admixed tumor specimens. Together, these findings underscore both the value and limitation of cell line models for unraveling tumor biology.

Multiple studies have explored genomic and transcriptomic markers for drug sensitivity in cancer cell lines [1, 3], but data on the proteome remain limited [11]. Our comparison of omics modalities for the identification of known drug-target gene or pathway relationships in CRC cell lines demonstrates the potential of global proteomic data to predict therapeutic responses. Consistent with our observation that DNA and mRNA measurements are poor predictors of protein abundance, protein level data outperformed mRNA, DNA copy number and mutation data in 11 out of 15 pairwise comparisons for five evaluated standard therapies (5-fluoruracil, SN-38, erlotinib a proxy for anti-EGFR antibody therapy, regorafenib, oxaliplatin). Furthermore, proteomic data more closely predicted known drug-target relationships, both at the individual gene and the target pathway levels. Pharmacological inhibition of targetable proteins associated with CRC cell line resistance or sensitivity to standard chemotherapies (5-FU and SN-38) identified markers that may be functionally implicated in determining drug responses, exhibiting synergy or antagonism in combination treatments, respectively. In addition, our data suggest that tumor proteomic subtypes may be useful predictors of drug responses, warranting further investigation in expanded studies. A caveat to our analysis is that we could not validate proteome-drug sensitivity relationships in our cohort of TCGA primary cancers due to insufficient cases with single-agent treatment and outcome data.

In summary, our integrative analysis demonstrates the utility of CRC cell lines as representative models of primary tumors at the proteome level, and highlights the potential of global proteomic data to inform personalized cancer medicine. Our data provide a rich resource for the scientific community and are available in public repositories and for interrogation via customized online research tools.

REFERENCES

1. **Garnett MJ**, **Edelman EJ**, **Heidorn SJ**, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570-5.
2. **Barretina J**, **Caponigro G**, **Stransky N**, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603-7.
3. **Iorio F**, **Knijnenburg TA**, **Vis DJ**, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016;166:740-54.
4. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-7.
5. **Guinney J**, **Dienstmann R**, **Wang X**, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350-6.
6. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513:382-7.
7. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 2006;6:813-23.
8. Chen B, Sirota M, Fan-Minogue H, et al. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. BMC Med Genomics 2015;8 Suppl 2:S5.
9. **Medico E**, **Russo M**, **Picco G**, et al. The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. Nat Commun 2015;6:7002.
10. Vincent KM, Findlay SD, Postovit LM. Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. Breast Cancer Res 2015;17:114.

11. Lawrence RT, Perez EM, Hernandez D, et al. The proteomic landscape of triple-negative breast cancer. Cell Rep 2015;11:630-44.

12. **Gholami AM**, **Hahne H**, **Wu Z,** et al. Global proteome analysis of the NCI-60 cell line panel. Cell Rep 2013;4:609-20.

13. Mouradov D, Sloggett C, Jorissen RN, et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. Cancer Res 2014;74:3238-47.

14. Wang J, Duncan D, Shi Z, et al. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 2013;41:W77-83.

15. Isella C, Terrasi A, Bellomo SE, et al. Stromal contribution to the colorectal cancer transcriptome. Nat Genet 2015;47:312-9.

16. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:8971.

17. Uhlen M, Oksvold P, Fagerberg L, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010;28:1248-50.

18. Poulogiannis G, Frayling IM, Arends MJ. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. Histopathology 2010;56:167-79.

19. Briggs S, Tomlinson I. Germline and somatic polymerase epsilon and delta mutations define a new class of hypermutated colorectal and endometrial cancers. J Pathol 2013;230:148-53.

20. Smyrk TC, Watson P, Kaul K, et al. Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. Cancer 2001;91:2417-22.

21. Radinsky R, Risin S, Fan D, et al. Level and function of epidermal growth factor receptor predict the metastatic potential of human colon carcinoma cells. Clin Cancer Res 1995;1:19-31.

22. **Wu Y**, **Peng Y**, Wu M, et al. Oncogene FOXK1 enhances invasion of colorectal carcinoma by inducing epithelial-mesenchymal transition. Oncotarget 2016.

23. **Xue C**, **Zhang Z**, **Yu H**, et al. Up-regulation of CNDP2 facilitates the proliferation of colon cancer. BMC Gastroenterol 2014;14:96.

24. MacFarlane AJ, Perry CA, McEntee MF, et al. Mthfd1 is a modifier of chemically induced intestinal carcinogenesis. Carcinogenesis 2011;32:427-33.

25. **Cheung HW**, **Cowley GS**, **Weir BA**, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc Natl Acad Sci U S A 2011;108:12372-7.

26. Moreb JS, Ucar-Bilyeu DA, Khan A. Use of retinoic acid/aldehyde dehydrogenase pathway as potential targeted therapy against cancer stem cells. Cancer Chemother Pharmacol 2017;79:295-301.

27. Bellamkonda K, Sime W, Sjolander A. The impact of inflammatory lipid mediators on colon cancer-initiating cells. Mol Carcinog 2015;54:1315-27.

28. Guo X, Xu B, Pandey S, et al. Disulfiram/copper complex inhibiting NFkappaB activity and potentiating cytotoxic effect of gemcitabine on colon and breast cancer cell lines. Cancer Lett 2010;290:104-13.

29. Wang W, McLeod HL, Cassidy J. Disulfiram-mediated inhibition of NF-kappaB activity enhances cytotoxicity of 5-fluorouracil in human colorectal cancer cell lines. Int J Cancer 2003;104:504-11.

30. Bavetsias V, Linardopoulos S. Aurora Kinase Inhibitors: Current Status and Outlook. Front Oncol 2015;5:278.

31. Cammareri P, Scopelliti A, Todaro M, et al. Aurora-a is essential for the tumorigenic capacity and chemoresistance of colorectal cancer stem cells. Cancer Res 2010;70:4655-65.

32. **Wu X**, **Liu W**, Cao Q, et al. Inhibition of Aurora B by CCT137690 sensitizes colorectal cells to radiotherapy. J Exp Clin Cancer Res 2014;33:13.

33. Ehrhardt H, Pannert L, Pfeiffer S, et al. Enhanced anti-tumour effects of Vinca alkaloids given separately from cytostatic therapies. Br J Pharmacol 2013;168:1558-69.

34. Calon A, Lonardo E, Berenguer-Llergo A, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat Genet 2015;47:320-9.

35. Longley DB, Harkin DP, Johnston PG. 5-fluorouracil: mechanisms of action and clinical strategies. Nat Rev Cancer 2003;3:330-8.

36. **Li H**, **Zhu F**, **Boardman LA**, et al. Aspirin Prevents Colorectal Cancer by Normalizing EGFR Expression. EBioMedicine 2015;2:447-455.

Author names in bold designate shared co-first authorship

FIGURE LEGENDS

Figure 1. Comparison of protein abundances between CRC cell lines and tumors. (**a**) Volcano plot indicating proteins overexpressed in cell lines (blue) or tumors (red) (FDR<5% and fold change>2); other genes are colored in grey. (**b**) The GO Biological Processes (BP) enriched for proteins overexpressed in cell lines (blue) or tumors (red) identified using WebGestalt [14]. (**c**) Overlap of stroma signatures with genes overexpressed in tumors *versus* other genes. *p* value for hypergeometric test. (**d**) Distributions of the signed -log10 *p* values (voom/limma) of the associations between protein abundance and tumor purity score for genes overexpressed in tumors *versus* other genes. *p* value for Wilcox rank sum test. (**e**) Heatmap of tumor stroma and epithelial protein marker expression in tumors and cell lines. The bar plot to the left of the heatmap represents the signed -log10 FDR (voom/limma) comparing protein abundances of tumor and cell line samples. (**f**) Box plots comparing protein abundance measurements for cell lines and tumors against tumor-cell specific IHC scores defined by the Human Protein Atlas. *p* values for Jonckheere's trend test.

Figure 2. Pathways associated with the hypermutation phenotype in CRC cell lines and tumors. (**a**) GSEA enrichment scores for significant KEGG pathways in cell lines and tumors. Red and blue bars represent the positively and negatively enriched pathways, respectively. The numbers in the parentheses represent the enriched FDR of the pathways. (**b**) Genes sorted by differential expression between hypermutated and non-hypermutated samples. Red and green represent overexpression in hypermutated and non-hypermutated samples, respectively. Bars in the bottom panel represent genes annotated to the mismatch repair pathway with blue bars

indicating the leading-edge genes reported by GSEA. (**c**) Comparison of protein abundance between hypermutated and non-hypermutated samples for the leading-edge genes identified from the cell line data.

Figure 3. Comparison of the correlations between mRNA and protein abundance in tumor and cell line data. (**a**) Correlations between steady state mRNA and protein abundance across genes within individual samples. (**b**) Correlations between mRNA and protein variation across cell line or tumor samples for each gene. (**c**) GSEA KEGG enrichment for average differences in mRNA-protein ranks across genes in both the cell line and tumor data. Genes colored in red are ranked higher in RNA, genes in green ranked higher in proteomics and blue are the leading-edge GSEA genes.

Figure 4. Comparison of cell lines and tumors to normal tissues based on protein abundance data. (**a**) Correlation of protein expression changes for cell line and tumor relative to normal tissue. (**b**) Overlap between up-regulated and down-regulated proteins (FDR<0.05, fold change>2) relative to normal. (**c**) Heat map showing protein expression in normal, tumor and cell line samples. (**d**) Coordinated protein expression changes within KEGG pathways determined using a linear mixed-effects model. Mean log fold change as compared to normal and heatmap of pathway expression shown for normal, tumor and cell line samples.

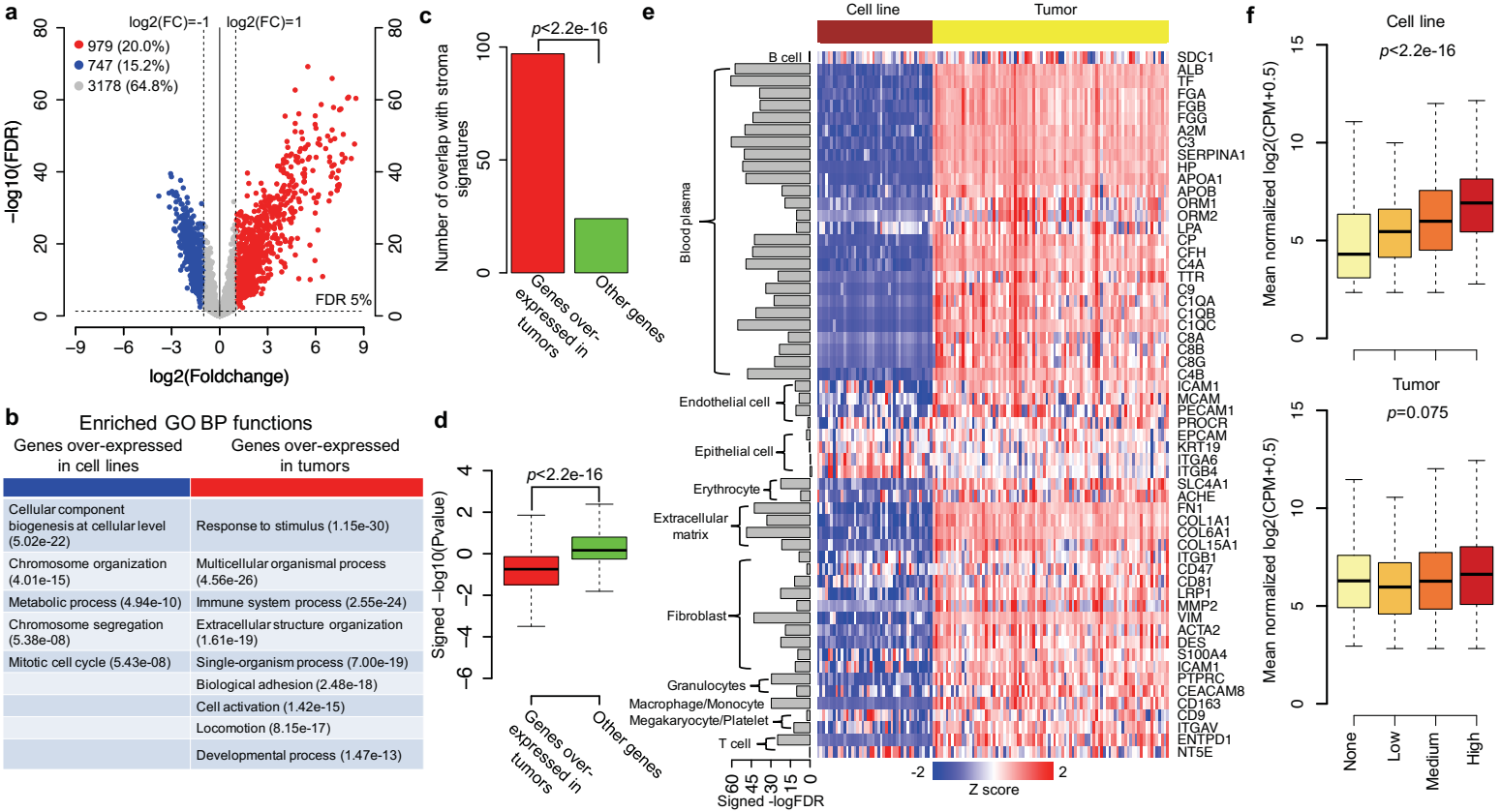
Figure 5. Proteome alterations associated with copy-number aberrations. (**a**) DNA copy-number spectra (% gain = red bars, % loss = blue bars, relative to ploidy) in cell lines and tumors. (**b**) Strengths of association for protein expression with corresponding DNA copy-number

changes (-log10(FDR)). Grey = not significantly associated with copy number alterations, blue = significant across proteomics cell line and tumor data only, green = significant for both proteomics and mRNA expression across cell line and tumor, red = candidate tumor suppressor and oncogenes.

Figure 6. Proteomics data utility for predicting therapeutic responses. (**a, b**) Associations of proteomic, mutation, DNA copy number and mRNA data with (**a**) established drug-target associations and (**b)** drug-pathway associations. Associations are shown for drug-target gene associations quantifiable at the protein level and significant in at least one of the four modalities as signed -log10(FDR) values from voom/limma and GSEA analyses, respectively. (**c**) Comparison of the utility of four omic modalities to predict drug sensitivity for 5-fluoruracil (5-FU), erlotinib, oxaliplatin, regorafenib and SN-38: proteomic data (red); RNA-Seq data (blue); CNA data (green); and exome mutation data (yellow). For each drug-omic modality combination, area under the receiver operating characteristic curve (AUROCs) were generated from 100 times of 5-fold cross-validations. The two-sided Wilcoxon rank sum test was used to compare the performance between protein-based models and models based on other omics data types. For each comparison, the *p* value is colored based on the color of the omic data type with significantly better performance. (**d-e**) Pharmacological targeting of proteins associated with resistance or sensitivity to (**d**) 5-FU or (**e**) SN-38. Bliss excess values are shown for drug combinations with 5-FU (at IC_{30} concentration) and SN-38 (at IC_{40} concentration) in HCT116 cells. The protein targets were restricted to those with FDR< 0.2 from the relevant voom/limma calculation; drugs are detailed in Supplementary Tables 4-5. *p*-values for Student's t-test. (**f**) Dose-response plots for selected compounds alone (black), with either a 5-FU or SN-38 (blue),

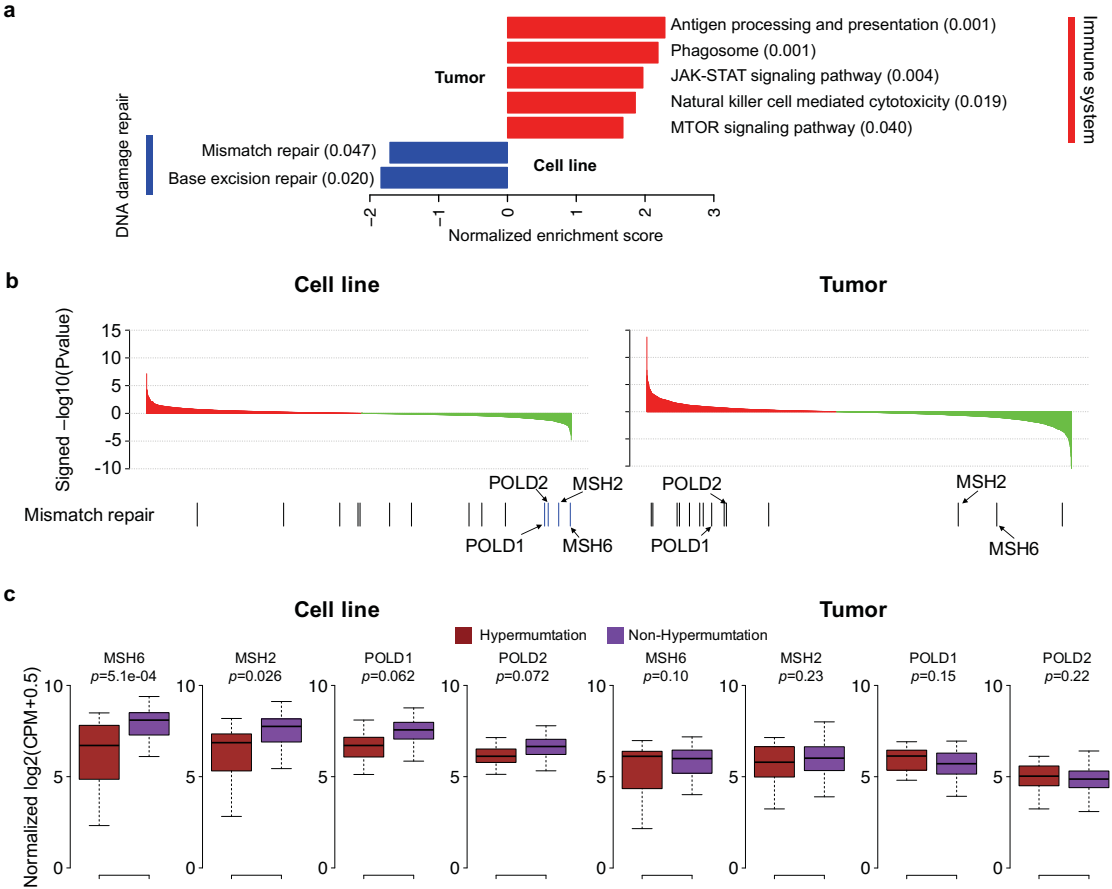
or the predicted response under the assumption of Bliss independence for the two compounds (green). Bliss synergy = blue line below green line; Bliss antagonism = blue line above green line.

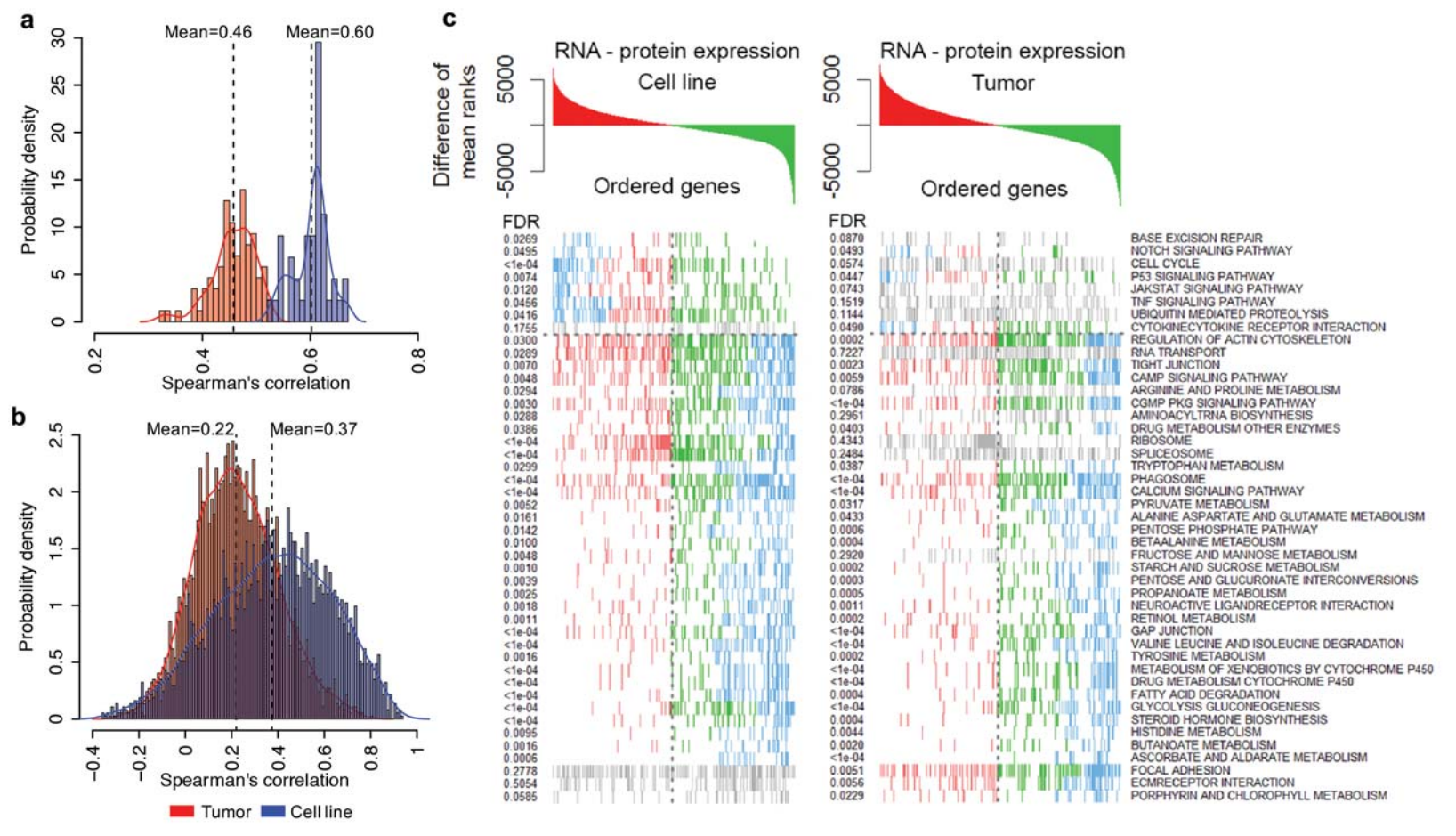
Figure 7. Concordance of proteomic CRC subtypes in cell lines and tumors. (**a**) Heatmap of protein abundances indicating proteomic subtypes for tumors (left panel) and cell lines (right panel). Samples are arranged along the X axis and genes are arranged along the Y axis. Increased expression (red) and decreased expression (blue) relative to the mean-centered and scaled expression of the gene (normalized CPM) across the samples. (**b**) Representation of genomic hallmarks among proteomic subtypes. (**c**) Drug responses of proteomic subtypes to 5-fluoruracil (5-FU), erlotinib, oxaliplatin, regorafenib and SN-38 treatment, and relationships with cell doubling time. Puni (univariate) is the P-value obtained from univariate ANOVA, and Padj (adjusted) is the P-value from two-way ANOVA adjusting for cell doubling time.

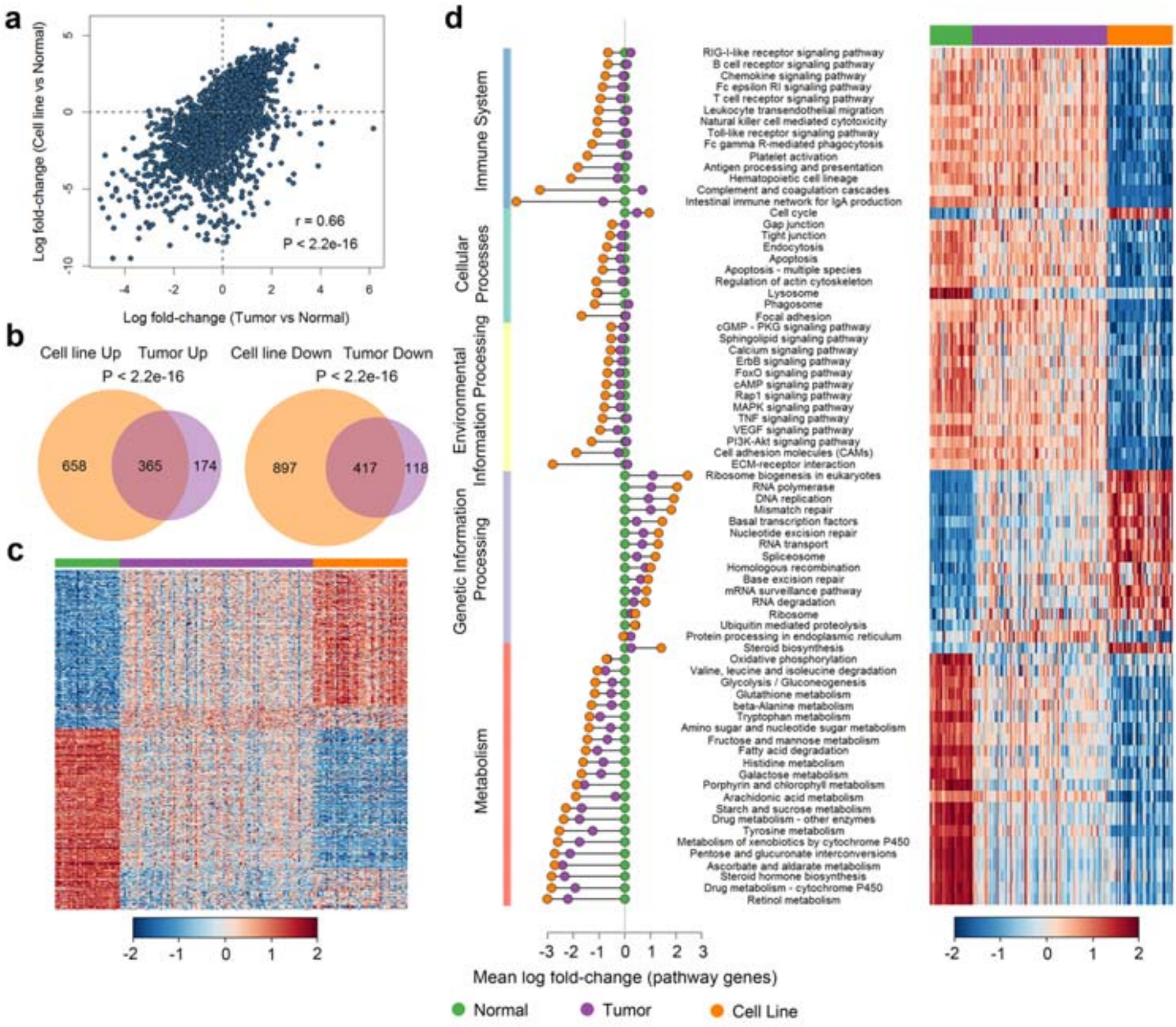


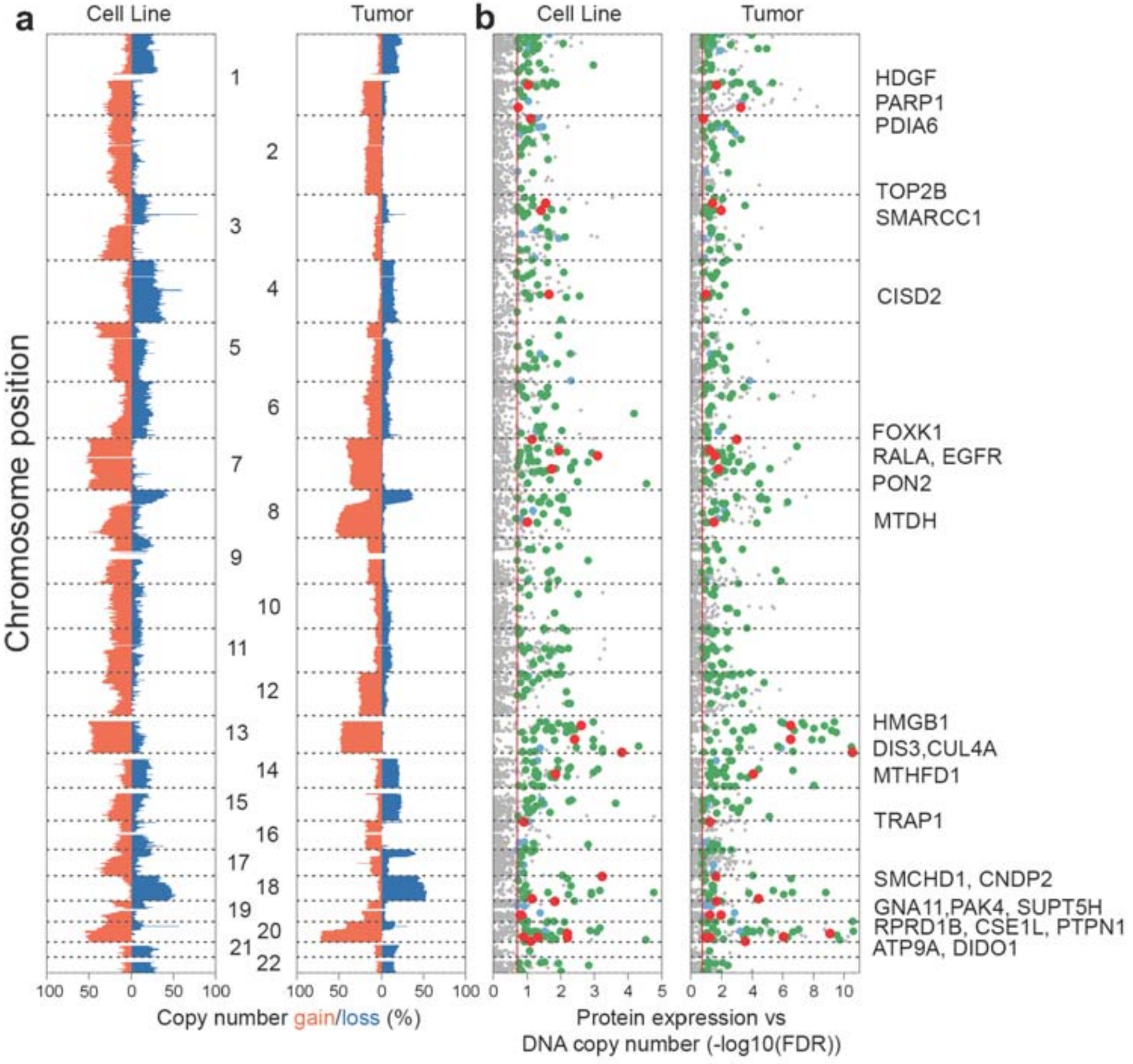
b Enriched GO BP functions

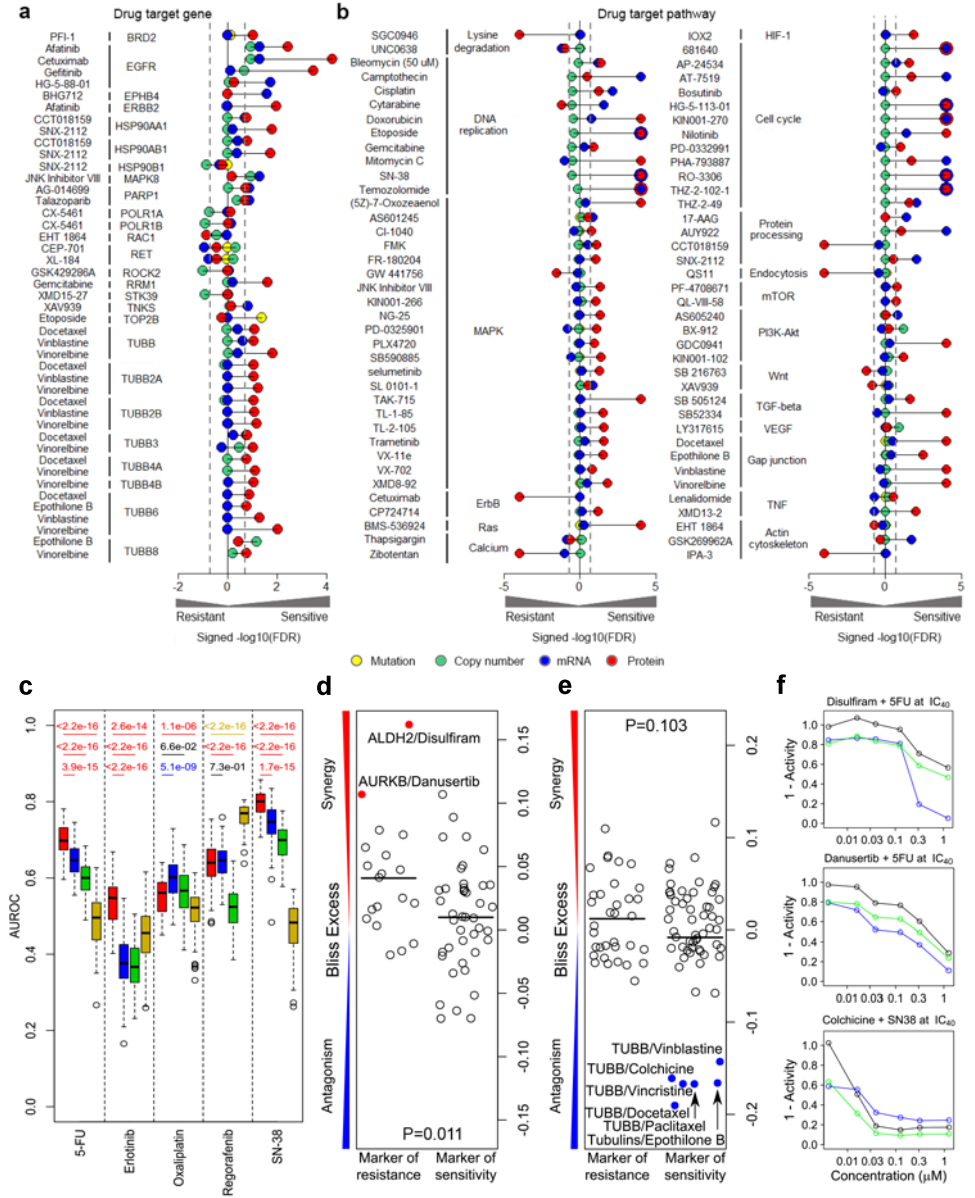
Genes over-expressed in cell lines	Genes over-expressed in tumors
Cellular component biogenesis at cellular level (5.02e-22)	Response to stimulus (1.15e-30)
Chromosome organization (4.01e-15)	Multicellular organismal process (4.56e-26)
Metabolic process (4.94e-10)	Immune system process (2.55e-24)
Chromosome segregation (5.38e-08)	Extracellular structure organization (1.61e-19)
Mitotic cell cycle (5.43e-08)	Single-organism process (7.00e-19)
	Biological adhesion (2.48e-18)
	Cell activation (1.42e-15)
	Locomotion (8.15e-17)
	Developmental process (1.47e-13)

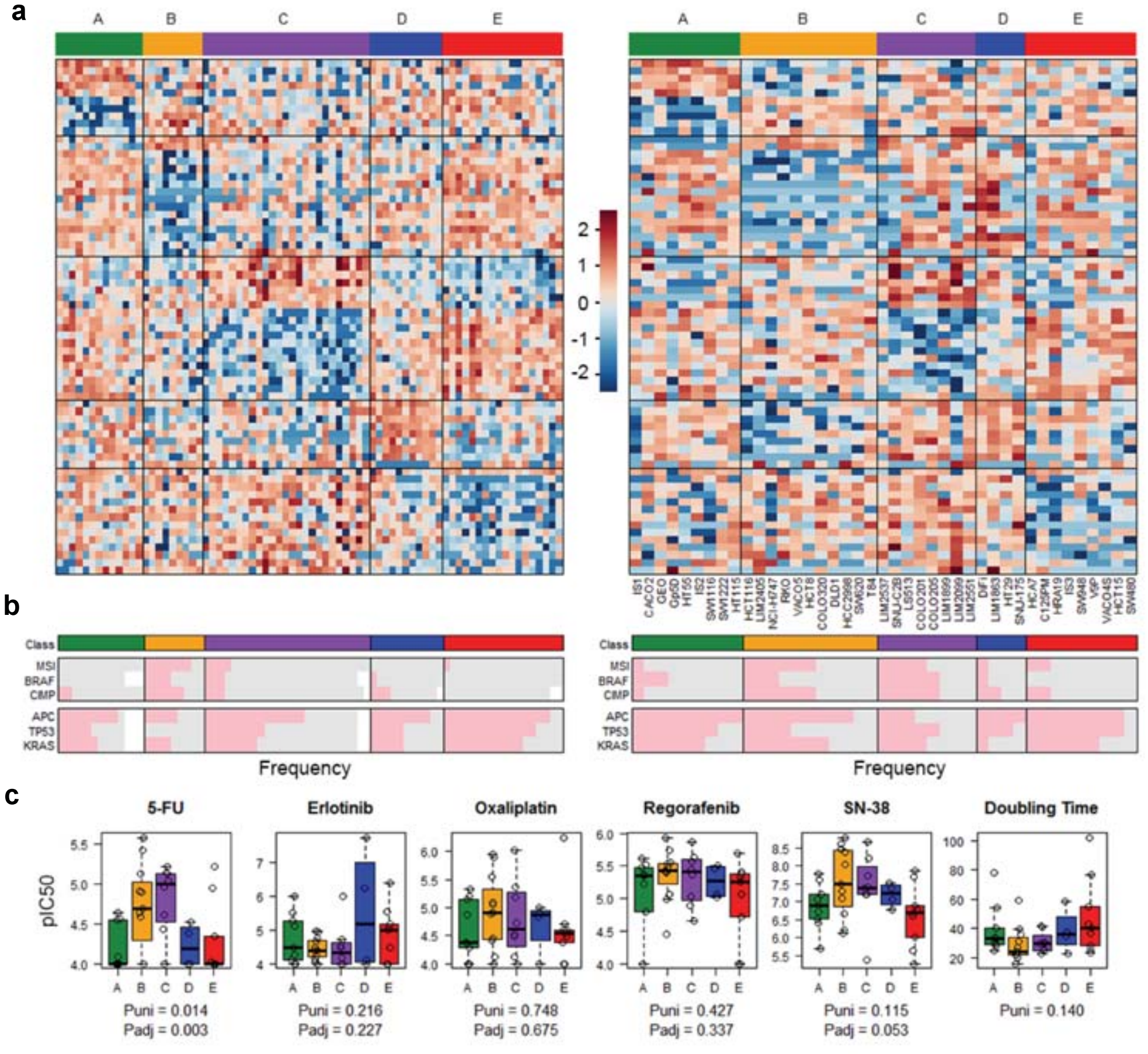

a
Antigen processing and presentation (0.001)
Phagosome (0.001)
JAK-STAT signaling pathway (0.004)
Natural killer cell mediated cytotoxicity (0.019)
MTOR signaling pathway (0.040)
Tumor
Immune system
DNA damage repair
Mismatch repair (0.047)
Base excision repair (0.020)
Cell line
−2
−1
0
1
2
3
Normalized enrichment score
b
Cell line
Tumor
Signed −log10(Pvalue)
15
10
5
0
-5
-10
POLD2
MSH2
POLD2
MSH2
Mismatch repair
POLD1
MSH6
POLD1
MSH6
c
Cell line
Tumor
Hypermumtation
Non-Hypermumtation
MSH6
p=5.1e-04
MSH2
p=0.026
POLD1
p=0.062
POLD2
p=0.072
MSH6
p=0.10
MSH2
p=0.23
POLD1
p=0.15
POLD2
p=0.22
Normalized log2(CPM+0.5)
10
5
0


a
Mean=0.46
Mean=0.60
Probability density
Spearman's correlation
b
Mean=0.22
Mean=0.37
Probability density
Spearman's correlation
Tumor
Cell line
c
RNA - protein expression
Cell line
Difference of mean ranks
5000
-5000
Ordered genes
RNA - protein expression
Tumor
5000
-5000
Ordered genes
FDR
0.0269
0.0495
<1e-04
0.0074
0.0120
0.0456
0.0416
0.1755
0.0300
0.0289
0.0070
0.0048
0.0294
0.0030
0.0288
0.0386
<1e-04
<1e-04
0.0299
<1e-04
<1e-04
0.0052
0.0161
0.0142
0.0100
0.0048
0.0010
0.0039
0.0025
0.0018
0.0011
<1e-04
<1e-04
0.0016
<1e-04
<1e-04
<1e-04
<1e-04
<1e-04
0.0095
0.0016
0.0006
0.2778
0.5054
0.0585
FDR
0.0870
0.0493
0.0574
0.0447
0.0743
0.1519
0.1144
0.0490
0.0002
0.7227
0.0023
0.0059
0.0786
<1e-04
0.2961
0.0403
0.4343
0.2484
0.0387
<1e-04
<1e-04
0.0317
0.0433
0.0006
0.0004
0.2920
0.0002
0.0003
0.0005
0.0011
0.0002
<1e-04
<1e-04
0.0002
<1e-04
<1e-04
0.0004
<1e-04
0.0004
0.0044
0.0020
<1e-04
0.0051
0.0056
0.0229
BASE EXCISION REPAIR
NOTCH SIGNALING PATHWAY
CELL CYCLE
P53 SIGNALING PATHWAY
JAKSTAT SIGNALING PATHWAY
TNF SIGNALING PATHWAY
UBIQUITIN MEDIATED PROTEOLYSIS
CYTOKINECYTOKINE RECEPTOR INTERACTION
REGULATION OF ACTIN CYTOSKELETON
RNA TRANSPORT
TIGHT JUNCTION
CAMP SIGNALING PATHWAY
ARGININE AND PROLINE METABOLISM
CGMP PKG SIGNALING PATHWAY
AMINOACYLTRNA BIOSYNTHESIS
DRUG METABOLISM OTHER ENZYMES
RIBOSOME
SPLICEOSOME
TRYPTOPHAN METABOLISM
PHAGOSOME
CALCIUM SIGNALING PATHWAY
PYRUVATE METABOLISM
ALANINE ASPARTATE AND GLUTAMATE METABOLISM
PENTOSE PHOSPHATE PATHWAY
BETAALANINE METABOLISM
FRUCTOSE AND MANNOSE METABOLISM
STARCH AND SUCROSE METABOLISM
PENTOSE AND GLUCURONATE INTERCONVERSIONS
PROPANOATE METABOLISM
NEUROACTIVE LIGANDRECEPTOR INTERACTION
RETINOL METABOLISM
GAP JUNCTION
VALINE LEUCINE AND ISOLEUCINE DEGRADATION
TYROSINE METABOLISM
METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
DRUG METABOLISM CYTOCHROME P450
FATTY ACID DEGRADATION
GLYCOLYSIS GLUCONEOGENESIS
STEROID HORMONE BIOSYNTHESIS
HISTIDINE METABOLISM
BUTANOATE METABOLISM
ASCORBATE AND ALDARATE METABOLISM
FOCAL ADHESION
ECMRECEPTOR INTERACTION
PORPHYRIN AND CHLOROPHYLL METABOLISM


a
Log fold-change (Cell line vs Normal)
Log fold-change (Tumor vs Normal)
r = 0.66
P < 2.2e-16
b
Cell line Up
Tumor Up
P < 2.2e-16
658
365
174
Cell line Down
Tumor Down
P < 2.2e-16
897
417
118
c
-2 -1 0 1 2
d
Immune System
Cellular Processes
Environmental Information Processing
Genetic Information Processing
Metabolism
RIG-I-like receptor signaling pathway
B cell receptor signaling pathway
Chemokine signaling pathway
Fc epsilon RI signaling pathway
T cell receptor signaling pathway
Leukocyte transendothelial migration
Natural killer cell mediated cytotoxicity
Toll-like receptor signaling pathway
Fc gamma R-mediated phagocytosis
Platelet activation
Antigen processing and presentation
Hematopoietic cell lineage
Complement and coagulation cascades
Intestinal immune network for IgA production
Cell cycle
Gap junction
Tight junction
Endocytosis
Apoptosis
Apoptosis - multiple species
Regulation of actin cytoskeleton
Lysosome
Phagosome
Focal adhesion
cGMP - PKG signaling pathway
Sphingolipid signaling pathway
Calcium signaling pathway
ErbB signaling pathway
FoxO signaling pathway
cAMP signaling pathway
Rap1 signaling pathway
MAPK signaling pathway
TNF signaling pathway
VEGF signaling pathway
PI3K-Akt signaling pathway
Cell adhesion molecules (CAMs)
ECM-receptor interaction
Ribosome biogenesis in eukaryotes
RNA polymerase
DNA replication
Mismatch repair
Basal transcription factors
Nucleotide excision repair
RNA transport
Spliceosome
Homologous recombination
Base excision repair
mRNA surveillance pathway
RNA degradation
Ribosome
Ubiquitin mediated proteolysis
Protein processing in endoplasmic reticulum
Steroid biosynthesis
Oxidative phosphorylation
Valine, leucine and isoleucine degradation
Glycolysis / Gluconeogenesis
Glutathione metabolism
beta-Alanine metabolism
Tryptophan metabolism
Amino sugar and nucleotide sugar metabolism
Fructose and mannose metabolism
Fatty acid degradation
Histidine metabolism
Galactose metabolism
Porphyrin and chlorophyll metabolism
Arachidonic acid metabolism
Starch and sucrose metabolism
Drug metabolism - other enzymes
Tyrosine metabolism
Metabolism of xenobiotics by cytochrome P450
Pentose and glucuronate interconversions
Ascorbate and aldarate metabolism
Steroid hormone biosynthesis
Drug metabolism - cytochrome P450
Retinol metabolism
-3 -2 -1 0 1 2 3
Mean log fold-change (pathway genes)
-2 -1 0 1 2
Normal
Tumor
Cell Line


a
Cell Line
Tumor
b
Cell Line
Tumor
Chromosome position
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
HDGF
PARP1
PDIA6
TOP2B
SMARCC1
CISD2
FOXK1
RALA, EGFR
PON2
MTDH
HMGB1
DIS3,CUL4A
MTHFD1
TRAP1
SMCHD1, CNDP2
GNA11,PAK4, SUPT5H
RPRD1B, CSE1L, PTPN1
ATP9A, DIDO1
100 50 0 50 100
100 50 0 50 100
Copy number gain/loss (%)
0 1 2 3 4 5
0 2 4 6 8 10
Protein expression vs
DNA copy number (-log10(FDR))


a
Drug target gene
PFI-1
Afatinib
Cetuximab
Gefitinib
HG-5-88-01
BHG712
Afatinib
CCT018159
SNX-2112
CCT018159
SNX-2112
SNX-2112
JNK Inhibitor VIII
AG-014699
Talazoparib
CX-5461
CX-5461
EHT 1864
CEP-701
XL-184
GSK429286A
Gemcitabine
XMD15-27
XAV939
Etoposide
Docetaxel
Vinblastine
Vinorelbine
Docetaxel
Vinblastine
Vinorelbine
Docetaxel
Vinblastine
Vinorelbine
Docetaxel
Vinorelbine
Docetaxel
Vinorelbine
Docetaxel
Epothilone B
Vinblastine
Vinorelbine
Epothilone B
Vinorelbine
BRD2
EGFR
EPHB4
ERBB2
HSP90AA1
HSP90AB1
HSP90B1
MAPK8
PARP1
POLR1A
POLR1B
RAC1
RET
ROCK2
RRM1
STK39
TNKS
TOP2B
TUBB
TUBB2A
TUBB2B
TUBB3
TUBB4A
TUBB4B
TUBB6
TUBB8
-2
0
2
4
Resistant
Sensitive
Signed -log10(FDR)
b
Drug target pathway
SGC0946
UNC0638
Bleomycin (50 uM)
Camptothecin
Cisplatin
Cytarabine
Doxorubicin
Etoposide
Gemcitabine
Mitomycin C
SN-38
Temozolomide
(5Z)-7-Oxozeaenol
AS601245
CI-1040
FMK
FR-180204
GW 441756
JNK Inhibitor VIII
KIN001-266
NG-25
PD-0325901
PLX4720
SB590885
selumetinib
SL 0101-1
TAK-715
TL-1-85
TL-2-105
Trametinib
VX-11e
VX-702
XMD8-92
Cetuximab
CP724714
BMS-536924
Thapsigargin
Zibotentan
Lysine degradation
DNA replication
MAPK
ErbB
Ras
Calcium
IOX2
681640
AP-24534
AT-7519
Bosutinib
HG-5-113-01
KIN001-270
Nilotinib
PD-0332991
PHA-793887
RO-3306
THZ-2-102-1
THZ-2-49
17-AAG
AUY922
CCT018159
SNX-2112
QS11
PF-4708671
QL-VIII-58
AS605240
BX-912
GDC0941
KIN001-102
SB 216763
XAV939
SB 505124
SB52334
LY317615
Docetaxel
Epothilone B
Vinblastine
Vinorelbine
Lenalidomide
XMD13-2
EHT 1864
GSK269962A
IPA-3
HIF-1
Cell cycle
Protein processing
Endocytosis
mTOR
PI3K-Akt
Wnt
TGF-beta
VEGF
Gap junction
TNF
Actin cytoskeleton
-5
0
5
Resistant
Sensitive
Signed -log10(FDR)
Mutation
Copy number
mRNA
Protein
c
AUROC
0.0
0.2
0.4
0.6
0.8
1.0
<2.2e-16
<2.2e-16
3.9e-15
2.6e-14
<2.2e-16
<2.2e-16
1.1e-06
6.6e-02
5.1e-09
<2.2e-16
<2.2e-16
7.3e-01
<2.2e-16
<2.2e-16
1.7e-15
5-FU
Erlotinib
Oxaliplatin
Regorafenib
SN-38
d
Synergy
Bliss Excess
Antagonism
ALDH2/Disulfiram
AURKB/Danusertib
P=0.011
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
Marker of resistance
Marker of sensitivity
e
Synergy
Bliss Excess
Antagonism
P=0.103
TUBB/Vinblastine
TUBB/Colchicine
TUBB/Vincristine
TUBB/Docetaxel
TUBB/Paclitaxel
Tubulins/Epothilone B
0.2
0.1
0.0
-0.1
-0.2
Marker of resistance
Marker of sensitivity
f
Disulfiram + 5FU at IC40
Danusertib + 5FU at IC40
Colchicine + SN38 at IC40
1 - Activity
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.03
0.1
0.3
1
Concentration (μM)


a
A
B
C
D
E
2
1
0
-1
-2
IS1
CACO2
GEO
Gp5D
HT55
IS2
SW1116
SW1222
HT115
HCT116
LIM2405
NCI-H747
RKO
VACO5
HCT8
COLO320
DLD1
HCC2998
SW620
T84
LIM2537
SNU-C2B
LS513
COLO201
COLO205
LIM1899
LIM2099
LIM2551
DiFi
LIM1863
HT29
SNU-175
HCA7
C125PM
HRA19
IS3
SW948
V9P
VACO4S
HCT15
SW480
b
Class
MSI
BRAF
CIMP
APC
TP53
KRAS
Frequency
c
5-FU
Erlotinib
Oxaliplatin
Regorafenib
SN-38
Doubling Time
pIC50
Puni = 0.014
Padj = 0.003
Puni = 0.216
Padj = 0.227
Puni = 0.748
Padj = 0.675
Puni = 0.427
Padj = 0.337
Puni = 0.115
Padj = 0.053
Puni = 0.140

Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity

Jing Wang, Dmitri Mouradov, Xiaojing Wang, Robert N. Jorissen, Matthew C. Chambers, Lisa J. Zimmerman, Suhas Vasaikar, Christopher G. Love, Shan Li, Kym Lowes, Karl-Johan Leuchowius, Helene Jousset, Janet Weinstock, Christopher Yau, John Mariadason, Zhiao Shi, Yuguan Ban, Xi Chen, Robert J. C. Coffey, Robbert J.C. Slebos, Antony W. Burgess, Daniel C. Liebler, Bing Zhang*, Oliver M. Sieber*

Supplementary File 1

This file contains Supplementary Data, Supplementary Methods and Supplementary Fig. 1-24.

Supplementary Data Set 1

This file contains Supplementary Tables 1-43.

SUPPLEMENTARY DATA

Proteomic detection of single nucleotide variants (SNVs) in CRC cell lines

WES and RNA-Seq captured a combined total of 111,022 nonsynonymous single nucleotide variants (nsSNVs), 19.4% of which were exclusively detected by RNA-Seq analysis with an enrichment of A:T to G:C transversions characteristic of RNA editing [1] (**Supplementary Fig. 4**). Of the detected nsSNVs 1,702 unique variants were observed at the proteomic level (**Supplementary Table 9**); 276 corresponded to somatic variants reported in the TCGA/COSMIC databases, and 952 were listed in the Single Nucleotide Polymorphism (dbSNP) database and are likely to be germline variants (**Supplementary Fig. 3a**). 678 SNVs were not captured in these databases, and these were significantly enriched in hypermutated as compared to non-hypermutated cell lines (p=9.7e-08, two-sided Wilcoxon rank-sum test), suggesting that most represented somatic changes (**Supplementary Fig. 3b**). As observed for known somatic variants, previously unreported SNVs had significantly higher predicted functional impact than the dbSNP-supported variants (**Supplementary Fig. 3c, Supplementary Table 9**). Non-dbSNP variants were associated with a stronger negative impact on protein abundance than dbSNP-supported variants (p<2.2e-16, two-sided Kolmogorov–Smirnov test), suggesting reduced protein stability or translational efficiency associated with these variants [2, 3].

The 276 TCGA/COSMIC-supported protein variants mapped to 248 genes, including 23 cancer genes in the Cancer Gene Census database such as KRAS, CTNNB1, TP53, EGFR, SF3B1, SMAD4, and CDH1. The list also included 27 targets of FDA-approved drugs or drugs in clinical trials [4], such as EGFR, ALDH1B1, HSD17B4, PARP4, GSR, MAP2K1, and

AKR1A1. Overall, we found TCGA/COSMIC-supported variants in protein drug-targets in 40 out of the 44 cell lines.

Proteomic, transcriptomic and mutational discordance among paired cell-lines

Included in our cell line panel were 2 pairs/triplets originally derived from the same tumor (COLO201/COLO205, DLD1/HCT8/HCT15) and 2 pairs/triplets derived from a primary tumor and metastatic derivatives (SW480/SW620, IS1/IS2/IS3). Assignments of proteomics subtypes exhibited some discordance for paired cell lines, with one outlier for the triplet of cell lines derived from the same tumor (DLD1/HCT8/HCT15) and one outlier each for the two primary-metastasis cell line pairs/triplets (SW480/SW620, IS1/IS2/IS3) (**Figure 7a**). CMS classifications were only confidently assigned for the IS1/IS2/IS3 triplet, but these also indicated discordance (**Supplementary Fig. 22a**). This proteomic and transcriptomic heterogeneity is consistent with heterogeneity observed at the genomic level between these paired cell lines, with 46, 372, 117, 116 and 129 mutational differences in the non-hypermutated pairs COLO201/COLO205, SW480/SW620, IS1/IS2, IS1/IS3 and IS2/IS3, and 4125, 2460 and 1584 for the hypermutated pairs DLD1/HCT8, DLD1/HCT15 and HCT8/HCT15.

SUPPLEMENTARY METHODS

Colorectal cancer cell lines and primary tumors. A total of 44 CRC cell lines were studied: C125, C135, C70, CACO2, COLO201, COLO205, COLO320-DM, DiFi, DLD1, GEO, Gp5D, HCA7, HCC2998, HCT116, HCT15, HCT8, HRA19, HT115, HT29, HT55, IS1, IS2, IS3, LIM1215, LIM1863, LIM1899, LIM2099, LIM2405, LIM2537, LIM2551, LS513, NCI-H747, RKO, SNU-175, SNU-C2B, SW1116, SW1222, SW480, SW620, SW948, T84, V9P, VACO4S, VACO5 (**Supplementary Table 1**). Cells were cultured in Dulbecco's modified Eagle's medium (DMEM; Gibco BRL Life Technologies) supplemented with 10% fetal bovine serum (FBS; Bovogen Biologicals), 100U/ ml penicillin-streptomycin (Sigma-Aldrich) at 37°C and 5% CO2 incubator. Cells were verified to be mycoplasma free using the Lookout Mycoplasma PCR Detection kit (Sigma-Aldrich). All the cell lines were authenticated at the Australian Genome Research Facility Ltd. (AGRF) (Parkville, VIC, Australia) by STR profiling analysis using the GenePrint 10 System (Promega). In addition, we retrieved previously published genomic, transcriptomic and proteomic data on 95 primary tumor specimens from 90 CRC patients and proteomics data from 60 normal colon biopsies from 30 patients from our original CPTAC study [2], as well as RNA-Seq data for 48 normal colon and rectum samples deposited by the TCGA (**Supplementary Table 2-3**). Fresh-frozen paired primary tumor and liver metastases samples from five patients with CRC were obtained from the Victorian Cancer Biobank, and gene expression microarray analysis performed using Affymetrix GeneChip® Human Gene 1.0 ST Arrays at the AGRF (accessible at the Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=wjodggyojjqjrgr&acc=GSE90814). This study was human research ethics committee-approved, and all patients gave informed consent.

LC/MS-MS. The protein extraction and tryptic digestion of the frozen cell line pellets were performed as previously described for TCGA CRC specimens [2]; however, the optimal cutting temperature (OCT) compound removal procedure was omitted, since OCT was not present. The resulting tryptic peptides were fractionated using off-line basic reversed phase high-pressure liquid chromatography (bRPLC). A total of 60 fractions were collected, concatenated, and analyzed on a Thermo Orbitrap-Velos mass spectrometer by reversed phase HPLC. All samples were analyzed on the same instrument system that was used for the TCGA CRC sample analysis and with the same chromatography components, separation conditions, instrument settings and laboratory personnel. Consistent with the TCGA CRC analysis, control samples from basal and luminal human breast carcinoma xenografts (CompRefs) were analyzed in alternating order after each set of five cell lines. Raw data were processed and used for database and spectral library searching using three different search engines, Myrimatch [5], Pepitome [6] and MS-GF+ [7], as previously described [2]. Protein assembly for the cell line data was performed using IDPicker 3 [8] at 0.2% PSM FDR and a minimum of 2 distinct spectra required per protein. To compare data from the cell line, tumor, and normal samples and to facilitate the integration between genomic and proteomic data, a gene-level assembly was performed for all cell line, tumor, and normal samples at 0.1% PSM FDR and a minimum of 2 distinct spectra required per protein. For the confidently identified proteins, we relaxed the PSM FDR threshold to 1% to rescue additional high quality PSMs that were excluded by the stringent PSM FDR threshold, as previously described [2]. For the 5 tumors and all 30 normal cases with proteomic measurements from duplicated samples, only the sample with a larger total spectral count was included for quantitative analyses. Raw data for the cell lines, database search results, and the two versions of assemblies can be found at the Mass spectrometry Interactive Virtual Environment (MassIVE,

ftp to massive.ucsd.edu, username: MSV000080374, password: a. HTTP access from the MassIVE website will be available after publication of the manuscript.)

Transcriptome sequencing. RNA samples from CRC cell lines were extracted from pellets collected at the same time as the samples for proteomics analysis using the AllPrep DNA/RNA Mini kit (Qiagen). Libraries were prepared for sequencing using the TruSeq Stranded Total RNA Library Preparation Kit (Illumina), pooled and clustered using the cBot system (Illumina) with TruSeq SR Cluster Kit v3 reagents (Illumina). Sequencing was performed on the Illumina HiSeq 2000 system with TruSeq SBS Kit v3 reagents (Illumina) at the AGRF. Each sample was sequenced to a depth of at least 28 million reads. Sequencing reads were quality assessed and trimmed for any remaining sequencing adaptor using Trimmomatic (v0.22) [9]; reads smaller than 50 bp were removed. Reads were subsequently aligned to human genome build Hg19 using Tophat (v2.0.8.Linux_x86_64) [10] with parameters -g 1, --library-type fr-firststrand. Gene level expression was quantified using Gencode v15 annotation using featureCounts with a parameter to account for stranded counting (-s 2) [11]. Data can be accessed at the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=ktadckourbkbbur&acc=GSE90830).

Exome-capture sequencing. Whole exome mutation data on 35 CRC cell lines from our cohort have been published previously [12]. Libraries for the additional DIFI, GEO, IS1, IS2, IS3, LIM1863, LIM2537, V9P and VAC05 cells were produced using the Nextera DNA Library Preparation Kit (Illumina), and 100bp paired-end read sequencing performed using the Nextera Rapid Capture Expanded Exome Enrichment Kit (Illumina) on an Illumina HiSeq 2000 System

at the AGRF. Sequence alignment and calling of SNVs and INDELs involved mapping with BWA (0.7.12) and variants calling with GATK (GATK-3.4-46). To ensure high-quality variant calling of putative somatic mutations in the absence of matched normal tissue by GATK, we created a hybrid pipeline between the GATK Germline and Somatic Best Practice Variant Detection Protocols (https://software.broadinstitute.org/gatk/best-practices/) as described previously [12]. Briefly, we aligned paired-end reads to the human reference genome (hg19) with BWA-mem followed by adding read groups, marking duplicates and re-ordering with Picard tools (1.69). We then carried out base quality score recalibration and INDEL realignment using GATK modules. Finally we applied the GATK variant caller 'HaplotypeCaller' and filtered reads using the 'VariantFiltration' module. The minimum Phred-scaled confidence threshold for calling variants was set to 30. The 'VariantFiltration' module excluded SNVs with: a quality by depth score (QD) <2.0, a Fisher strand score (FS) >60.0, Mapping Quality Rank Sum Test (MappingQualityRankSum) < -12.5 and relative positioning of ALT alleles within reads (ReadPosRankSum) < -8.0. For INDELs, the following exclusion criteria were used: QD <2.0, FS >200.0 and ReadPosRankSum < -20.0. These filters ensured: (1) high confidence variant calls based on unfiltered depth of non-reference samples (QD); (2) low strand bias for detection of variants (FS) as strand bias is indicative of false positive calls; (3) checks for similar mapping qualities between REF and ALT alleles (MappingQualityRankSum) and checks to determine whether there was a position bias within the reads between ALT and REF alleles – ALT (but not REF) alleles occurring at end of reads is indicative of false positive calls. To remove putative germline variants in the absence of matched normal data we annotated detected alterations against databases of human germline variations including the Single Nucleotide Polymorphism database (dbSNP, build 135, SAO = 1), 1000 Genomes Project database (build

20110521), Mills *et al*. data set of small insertions and deletions [13] and germline variants detected in 114 normal colorectal tissues analyzed in our laboratories. Regions of known germline chromosomal segmental duplications were excluded to reduce the possibility of false-positive variants caused by read mismapping [14]. For analyses of cancer gene mutations, following variant classification was considered: “FRAME SHIFT” / “Frame Shift Del” / “Frame Shift Ins", "CODON DELETION" / "In Frame Del", "CODON INSERTION" / "In Frame Ins", "SPLICE SITE ACCEPTOR" / "SPLICE SITE DONOR" / "Splice Site", "STOP GAINED" / "Nonsense Mutation", “NON SYNONYMOUS CODING" / "Missense Mutation", "CODON CHANGE PLUS CODON DELETION", "CODON CHANGE PLUS CODON INSERTION", "STOP LOST", "START GAINED", "START LOST”.

Variant peptide identification and analysis. To identify variant peptides, we used a customized protein sequence database approach [15, 16] wherein we derived customized protein sequence databases from matched RNA-seq data and then performed database searches using the customized databases for individual samples. Sequence alignment and calling of SNVs and INDELs involved mapping with STAR 2.5.0c and variants calling with GATK (3.5-0-g36282e4). To ensure high-quality variant calling by GATK, we followed the GATK Best Practice Variant Detection protocol on RNA-Seq (http://gatkforums.broadinstitute.org/dsde/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail). Briefly, we aligned reads to the human reference genome (hg19) with STAR followed by adding read group, marking duplicates and re-ordering with Picard tools (1.78). We then applied the GATK pipeline that includes modules ‘SplitNCigarReads’, ‘HaplotypeCaller’ and ‘VariantFiltration’. The minimum Phred-scaled

confidence threshold for calling variants was setting to 20. The 'VariantFiltration' module excluded SNVs with: a quality by depth score (QD) <2.0, a Fisher strand score (FS) >30.0 or clusters of at least 3 SNPs that were within a window of 35 bases between them. These filters ensured: (1) high confidence variant calls based on unfiltered depth of non-reference samples (QD); (2) low strand bias for detection of variants (FS)—as strand bias is indicative of false positive calls; (3) filtering of many false variant calls introduced by mapping error of RNA-Seq reads. For customized database construction and variant peptide identification we used the R package *customProDB* [15] to annotate variations predicted from RNA-seq, including mapping to dbSNP138 and COSMIC66 databases. For each sample, *customProDB* generates a protein FASTA database by appending proteins with nonsynonymous protein coding SNVs and aberrant proteins to the end of the standard RefSeq human protein sequence database. Peptide identification was performed for each sample separately using corresponding customized FASTA database and MS-GF+ and MyriMatch 2.1.87. Search settings were identical to those described above. IDPicker 3 was used for protein assembly as described earlier, except that the data set was filtered at 1% PSM FDR and a minimum of 5 spectra identified per protein. The full data set consisted of 9,983 protein groups with 4.3% protein FDR**.** Identified SNVs were further annotated for existence in the somatic variant list published by TCGA [17] (*i.e.*, TCGA-somatic variants), existence in the COSMIC66 database (that is, COSMIC-supported variants), and existence in the dbSNP138 database (*i.e.*, dbSNP-supported variants). Functional impact of the SNVs was analyzed using MutationAssessor [18] and Sorting Intolerant From Tolerant (SIFT) [19].

SNP microarray analysis. SNP array data on 38 cell lines from our cohort have been published previously [12]. SNP array assays on the additional DiFi, GEO, IS1, IS2, IS3 and V9P cells were performed at the AGRF using CytoSNP-850K v1.1 BeadChips (Illumina). SNP array data were

processed using GenomeStudio software (Illumina), and SNPs with detected copy-number variation in a reference set of 637 normal tissue samples were excluded from downstream analysis as described previously [12]. The median call rate for the cell line samples was 97.3% (range 93.5-99.7%). DNA copy number segmentation with adjustment for normal contamination and intra-tumor heterogeneity was performed using the OncoSNP v2.18 suite [20], and the proportion of samples with gain or loss relative to ploidy (modal chromosome copy number) quantified at the SNP level. Average mean log R ratios were calculated for genes from the segmented data based on their RefSeq genomic positions.

VOOM/LIMMA analysis. The application of Voom to count data [21] assumes that the gene-wise mean-variance relationship should be smoothly decreasing with the count size. This assumption was met by filtering for quantifiable proteins or mRNAs, defined as CPM>20 in 20% of samples for protein spectral counts and CPM>1 in 20% of samples for RNA-Seq counts (**Supplementary Fig. 24**). Then, based on the quantifiable proteins or mRNAs, we used voom to normalize the proteomics or RNA-Seq data and performed differential gene expression analyses utilizing limma [22]. Voom/limma analyses were performed using Limma [22] and edgeR [23] R packages, and method sensitivity and specificity for spectral count data were validated using the spike-in data set generated by the 2015 study of the Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (ABRF) (ftp://iprg_study@ftp.peptideatlas.org/ (password ABRF329)) (**Supplementary Fig. 1**). Briefly, the 2015 iPRG study was based on four artificially made samples of known composition, each containing a constant background 200 ng of tryptic digests of *S.cerevisiae* (ATCC strain 204508/S288c). Each sample was separately spiked with different quantities of six individual protein digests and analyzed in triplicate by LC-

MS/MS acquisitions (total of 12 runs) using a Thermo Scientific Q-Exactive mass spectrometer. Data were acquired in data-dependent (DDA) mode. The MS/MS spectra were searched against the provided target-decoy protein database using three sequence search engines, OMSSA [24], MS-GF+ [7] and Comet [25]. The search results were first validated at the peptide-spectrum match (PSM) level by PeptideProphet [26], employing decoy-assisted semi-parametric modeling [27]. The results from the three search engines were combined using iProphet [28]. The LC-MS features were identified and quantified with Skyline [29] v.2.6.0.6851. The original Skyline-based quantification in a tab-delimited table form was downloaded from the ftp site. Voom/limma identified spike-in samples with a sensitivity of 87.1%, specificity of 99.9% positive predictive value of 93.1% and negative predictive value of 99.9% based on FDR<0.05 and greater than 2-fold change (**Supplementary Fig. 1a-b**), and voom/limma estimated fold-changes were highly correlated with expected fold-changes (Spearman's correlation=0.95, p<2.2e-16, **Supplementary Fig. 1c**).

Human Protein Atlas. The Human Protein Atlas data were downloaded from http://www.proteinatlas.org/about/download (cancer.csv and proteinatlas.tab), which contained the IHC expression scores of 185,406 patients for 16,235 proteins on colorectal tumor samples. Data were filtered for antibodies with "supportive" evidence. Summary tumor protein scores were classified as not detected, low, medium or high staining groups based on the mode of respective individual sample scores.

Tumor stroma markers. Markers for tumor cells and stroma components were assembled from key human cell phenotype markers (BD Human and Mouse CD Marker Handbook,

https://www.bdbiosciences.com/documents/cd_marker_handbook.pdf; http://www.biolegend.com/cell_markers; [30]), blood group systems [31], blood plasma [32] and extracellular matrix components [33] (**Supplementary Table 41**)

KEGG pathway. The KEGG pathways and corresponding annotations were downloaded using KEGG API (http://www.kegg.jp/kegg/rest/keggapi.html) [34]. We only considered the pathways from the classes "Metabolism", "Genetic Information Processing", "Environmental Information Processing", "Cellular Processes" and "Organismal Systems", which contained 229 pathways and 6,488 unique annotated genes.

Correlation between steady state mRNA and protein abundance. Because steady state comparisons require mRNA and protein measurements within a sample to be comparable, we used FPKM (Fragments Per Kilobase Million) and NSAF (Normalized Spectral Abundance Factor) to normalize the RNA-Seq and proteomics data of tumors and cell lines. Then, based on the 8,874 overlapping genes among the four data sets, we calculated the Spearman's correlation coefficients between FPKM and NSAF measurements for both tumors and cell lines.

Correlation between mRNA and protein variation. To evaluate mRNA and protein variations across samples, we focused on 3,718 overlapping quantifiable genes identified from Voom among RNA-Seq and proteomics data of tumors and cell lines. The different sample sizes of tumor and cell line cohorts may cause the correlations between mRNA and protein variation from these two data sets to be incomparable. Thus, for the tumor data, we randomly selected 44 samples and calculated the Spearman's correlations

between mRNA and protein variations across these 44 samples. We repeated this process 100 times and calculated the mean correlation for each gene. For the cell line data, we directly calculated the Spearman's correlations between mRNA and protein variations across the 44 cell line samples. Then, based on the Spearman's correlation of the tumor and cell line data, we identified the enriched KEGG pathways based on the two-sided Kolmogorov-Smirnov test under FDR 5%.

Correlation of relative mRNA-protein abundances. To identify pathways that are modulated at the post-transcriptional level in cell lines and tumors, we used the FPKM and NSAF normalized data of the 8,874 overlapping genes among the four protein and mRNA data sets. We calculated the mean differences between mRNA and protein ranks for each gene within individual samples across each cohort and then performed the GSEA enrichment analysis against KEGG pathways (excluding the overview pathways in the "Metabolism" and "Organismal systems" classes) to identify the enriched pathways under a 5% FDR.

Pathway signature identification. To assess whether genes in a given KEGG pathway have differing expression in tumors or cell lines relative to normal colorectal tissue, we modelled the protein or mRNA expression levels (cpm values for quantifiable genes) of pathway members using a linear mixed-effects model (lme4 R package, [35]). Genes and sample type were treated as fixed effects (each as categorical variable), and the interaction terms between the genes and sample type (grouped by sample type) as random effects. The coefficient for sample type was interpreted as an aggregate measure of expression change for the pathway proteins/mRNAs in tumors or cell lines relative to normal tissue. P values were calculated using the degrees of

freedom for the sample type coefficients as the number of respective pathway proteins or mRNAs minus one.

Comparison of the impact of copy number alteration on protein abundance for cell lines and tumors. Evaluation of the association between copy number alteration and protein or mRNA levels was carried out for genes with complete gene-level log R ratio data (not all gene level data could be computed due to probe failure, see call rates) and which had quantifiable expression from Voom: 4,878 proteins and 12,277 mRNAs for cell lines, and 4,344 proteins and 13,269 mRNAs for tumors. We performed voom/limma analysis utilizing robust linear regression for gene-level log R ratios against protein or RNA-Seq expression levels. Voom/limma analysis was run for each gene-level log R ratio state across all genes, retrieving only the relevant FDR adjusted statistic of the gene in question. Results were aggregated and overlapping significant associations identified between protein and mRNA data for cell lines and tumors.

Comparison of the effect of candidate oncogene-targeting shRNAs on the proliferation of colon cancer cell lines. The shRNA gene level data was downloaded from the Achilles project website (https://portals.broadinstitute.org/achilles/datasets/5/download) and contained eight colon cancer cell lines overlapped with our 44 cell lines. We calculated the spearman's correlation between shRNA score and log-transformed copy number data across eight cell lines for each candidate oncogene. The negative correlation indicates the gene knockdown affects the cell proliferation. Because of the limited sample size, we identified the significant candidate

oncogenes based on r<-0.5 instead of the p value. If one gene has multiple shRNAs, this gene was selected only if all shRNAs were concordant.

Drug sensitivity studies. Cells were seeded into 384-well plates (1000 cells/well) in DME (Gibco) supplemented with 10% FBS (Bovogen Biologicals). Oxaliplatin (Cat# S1224), erlotinib (Cat# S7786) and regorafenib (Cat# S1178) were purchased from Selleck Chemicals. 5-fluorouracil (5-FU) (Cat# F6627) was obtained from Sigma. Compounds were titrated in DMSO (10-point 3-fold dilution series) and added to the cells in quadruplicate using liquid handling robotics. Final DMSO concentration in all wells was 0.25%. After incubation with compounds for 72hr, cell viability was determined using CellTiter-Glo-2 (Promega) according to manufacturer's instructions and calculated as a percentage of DMSO (100%) and 1uM bortezomib (0%, Cat# S1013, Selleck Chemicals). Data was analyzed in Pipeline Pilot (BIOVIA) and the IC_{50} values calculated using a four-parameter logistic nonlinear regression model. Data were summarized as pIC50 +/- SD= with 2-3 independent experiments for each cell line. For drug combination screening in HCT116 cells, 123 drugs were accessed from Compounds Australia, Griffith University, Australia (**Supplementary Tables 4-5**). For each compound at a given dose, we calculated the Bliss excess as $BE = f_{combo} - f_{single} - f_{2nd\ drug} + (f_{single} \times f_{2nd\ drug})$ for duplicate experiments. We then calculated the average BE over the doses for that compound in combination with either 5-FU or SN38.

GDSC data. GDSC (Genomics of Drug Sensitivity in Cancer) drug sensitivity data [36] were downloaded from http://www.cancerrxgene.org/downloads (version 07/04/2016), which

contained 18 colon cancer cell lines used in this paper and 251 drugs. Because the sensitivity data of some drugs were missed in many of the 18 cell lines, we finally kept 210 drugs with at most five missing values for analysis. Drug-target gene and KEGG pathway level relationships tested are summarized in **Supplementary Table 34**.

Comparison of omic modalities for prediction of drug sensitivity. To compare the utility of the proteomics, RNA-Seq, CNA and mutation data for predicting drug sensitivity to 5-fluorouracil, erlotinib, oxaliplatin, regorafenib and SN-38 in our 44 CRC cell line panel, 3269 common genes assayed by all four omics platforms were used as the features for the prediction. Following the approach of Haibe-Kains *et. al.* [37], the 44 cell lines were dichotomized into sensitive and resistant groups for each drug based on the median of their respective pIC50 values. For each drug-omics modality combination, random forests models were constructed and evaluated using 100 times of 5-fold cross-validation based on AUROC (area under the receiver operating characteristic curve). During the training phase of each cross-validation, we used 1000 trees and optimized the number of features randomly sampled as candidates at each split from a grid of 100 pre-defined numbers using an inner-loop cross-validation. To compare the performance between proteomics data and other omics data, the two-sided Wilcoxon rank sum test was performed.

Cell line proteomic and CMS subtype predictions. To assign cell lines to our previously identified proteomic subtypes [2], normalized cpm data from voom were transformed into z-scores and the R package *pamr* (http://CRAN.R-project.org/package=pamr) was used to apply our predefined signature genes to the cell line scaled expression matrix. To identify the optimal value of the shrinkage parameter for our PAM prediction model, we selected the value that minimized leave-one-out cross-validated misclassification error for 79 tumor samples (error rate<2%). We

assigned five proteomics subtypes to 40 cell lines with probability more than 0.8 using 1,377 signature genes. For each sample in a subtype, we calculated the Pearson correlation with others in the same subtype. To assign CMS subtypes to cell lines and a dataset of 5 matched primary and metastatic tumors pairs, we used the CMSclassifier package in R (https://github.com/Sage-Bionetworks/CMSclassifier). To make the RNA-Seq and microarray data compatible with the microarray data background set supplied with the package, gene expression values were quantile normalized to the reference distribution. Class assignments were made based on consistent predictions from both the nearest random forest and single sample predictors.

Resources. To make the quantitative data described in this paper available to the scientific community, we developed a web application CRCOmics (http://crcomics.zhang-lab.org), which allows users to perform differential, correlation, and pathway analyses to compare cell lines and tumors, and to visualize analysis results using various types of statistical plots. To enable visualization of the variant peptides identified in this study in the context of the human genome, we converted PSMs from the customized search results into the *proBAM* format [38], which can be download or accessed in a JBrowse-based genome browser (http://proteogenomics.zhang-lab.org/).

SUPPLEMENTARY REFERENCES

1. Gerber AP, Keller W. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem Sci 2001;26:376-84.
2. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513:382-7.
3. Halvey PJ, Wang X, Wang J, et al. Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. Cancer Res 2014;74:387-97.
4. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519-25.
5. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 2007;6:654-61.
6. Dasari S, Chambers MC, Martinez MA, et al. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. J Proteome Res 2012;11:1686-95.
7. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 2014;5:5277.
8. Holman JD, Ma ZQ, Tabb DL. Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker. Curr Protoc Bioinformatics 2012;Chapter 13:Unit13 17.
9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-20.

10. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562-78.
11. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923-30.
12. Mouradov D, Sloggett C, Jorissen RN, et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. Cancer Res 2014;74:3238-47.
13. Mills RE, Pittard WS, Mullaney JM, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res 2011;21:830-9.
14. Bailey JA, Gu Z, Clark RA, et al. Recent segmental duplications in the human genome. Science 2002;297:1003-7.
15. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 2013;29:3235-7.
16. Wang X, Slebos RJ, Wang D, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res 2012;11:1009-17.
17. TCGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-7.
18. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118.
19. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073-81.

20. Yau C, Mouradov D, Jorissen RN, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome biology 2010;11:R92.

21. Law CW, Chen Y, Shi W, et al. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:R29.

22. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.

23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139-40.

24. Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. J Proteome Res 2004;3:958-64.

25. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics 2013;13:22-4.

26. Keller A, Nesvizhskii AI, Kolker E, et al. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383-92.

27. Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res 2008;7:286-92.

28. Shteynberg D, Deutsch EW, Lam H, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 2011;10:M111 007690.

29. MacLean B, Tomazela DM, Shulman N, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 2010;26:966-8.

30. Kalluri R, Zeisberg M. Fibroblasts in cancer. Nat Rev Cancer 2006;6:392-401.

31. Mitra R, Mishra N, Rath GP. Blood groups systems. Indian J Anaesth 2014;58:524-8.

32. Schenk S, Schoenhals GJ, de Souza G, et al. A high confidence, manually validated human blood plasma protein reference set. BMC Med Genomics 2008;1:41.

33. Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. Nat Rev Mol Cell Biol 2014;15:786-801.

34. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40:D109-14.

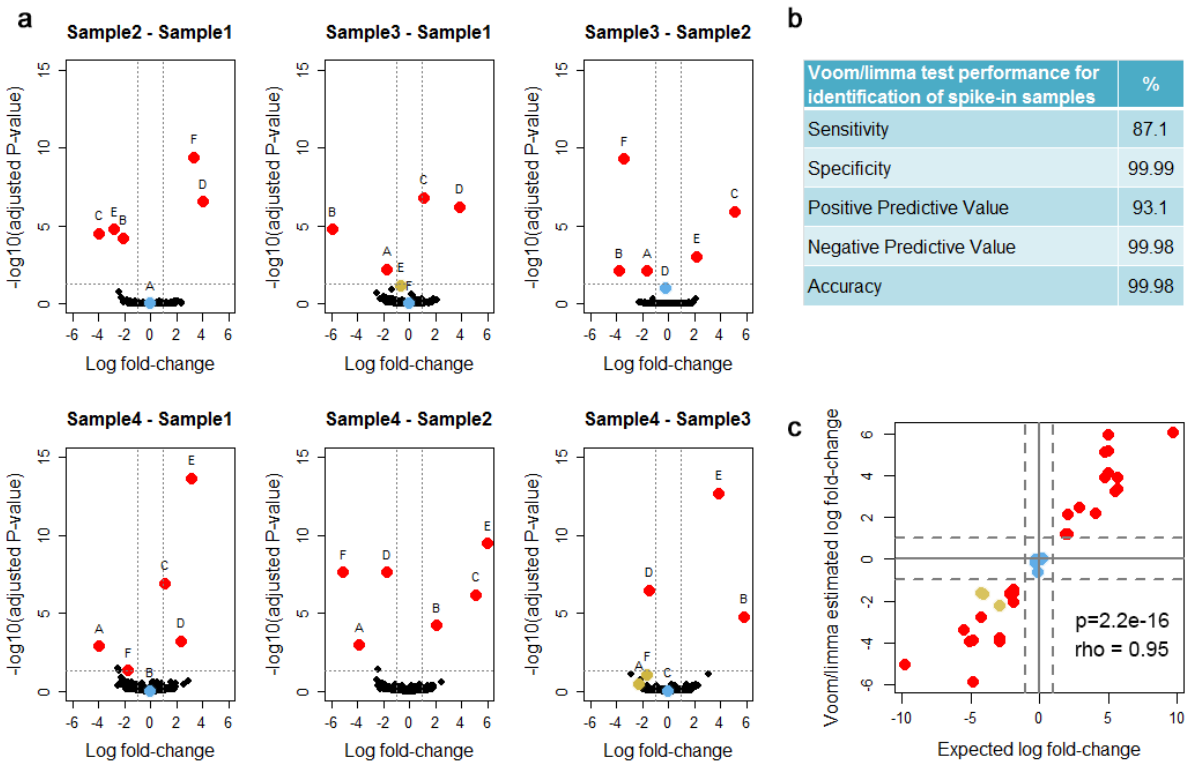
35. Bates D, Maechler M, Bolker B, et al. Fitting Linear Mixed-Effects Models Using lme4. J Statistical Software 2015;67:1-48.

36. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016;166:740-54.

37. Haibe-Kains B, El-Hachem N, Birkbak NJ, et al. Inconsistency in large pharmacogenomic studies. Nature 2013;504:389-93.

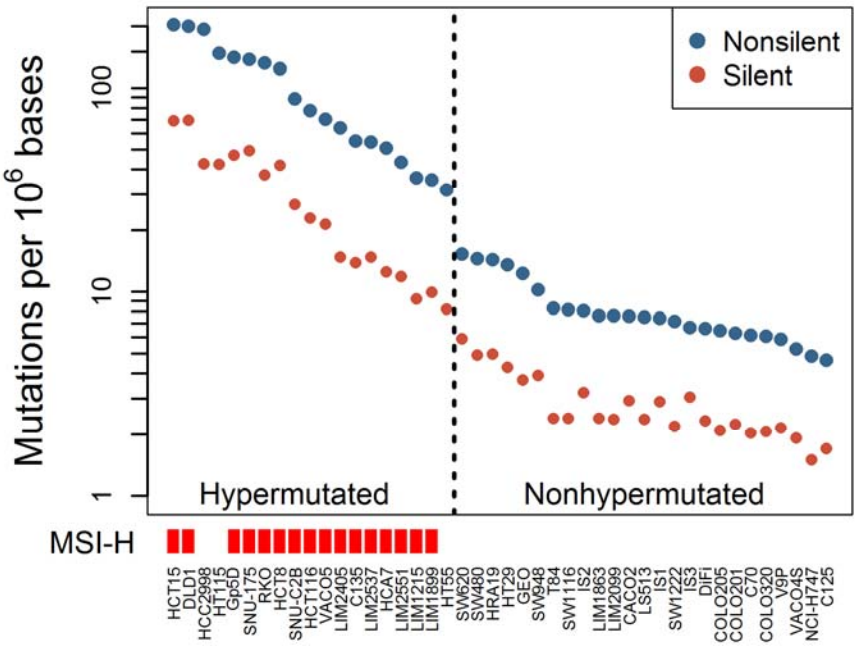
38. Wang X, Slebos RJ, Chambers MC, et al. proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data. Mol Cell Proteomics 2016;15:1164-75.

SUPPLEMENTARY FIGURES

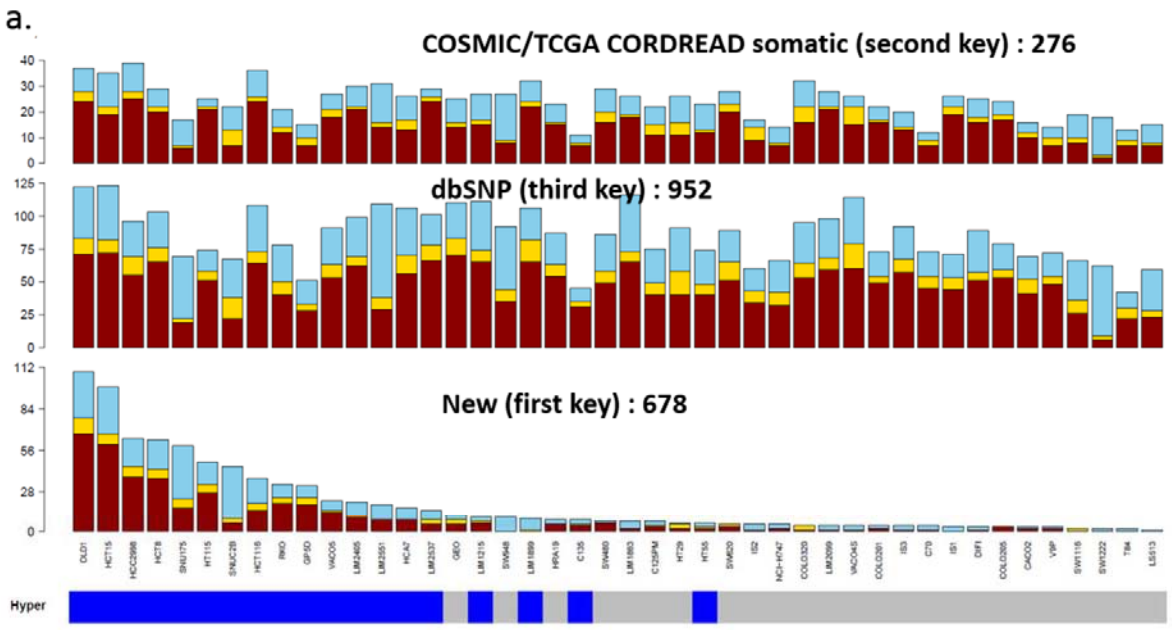


Voom/limma test performance for identification of spike-in samples	%
Sensitivity	87.1
Specificity	99.99
Positive Predictive Value	93.1
Negative Predictive Value	99.98
Accuracy	99.98

Supplementary Fig. 1. Voom/limma method sensitivity and specificity for protein spectral count data using spike-in data. (**a**) Six spike-in sample comparisons showing identified true positives (red), true engatives (blue) and false negatives (gold). (**b**) Test performance of voom/limma across spike-in experiments. (**c**) Comparison of expected and voom/limma derived log fold-changes for spike-in proteins.

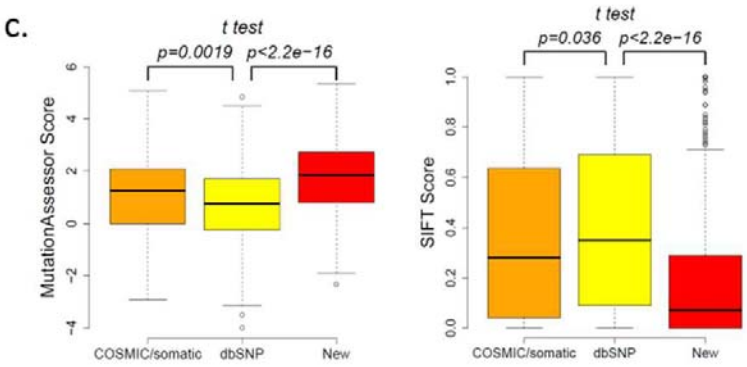


Supplementary Fig 2. Mutation frequencies in 44 human CRC cell lines. Cell lines segregated into distinct hypermutated and non-hypermutated cases. MSI-H, microsatellite instability-high



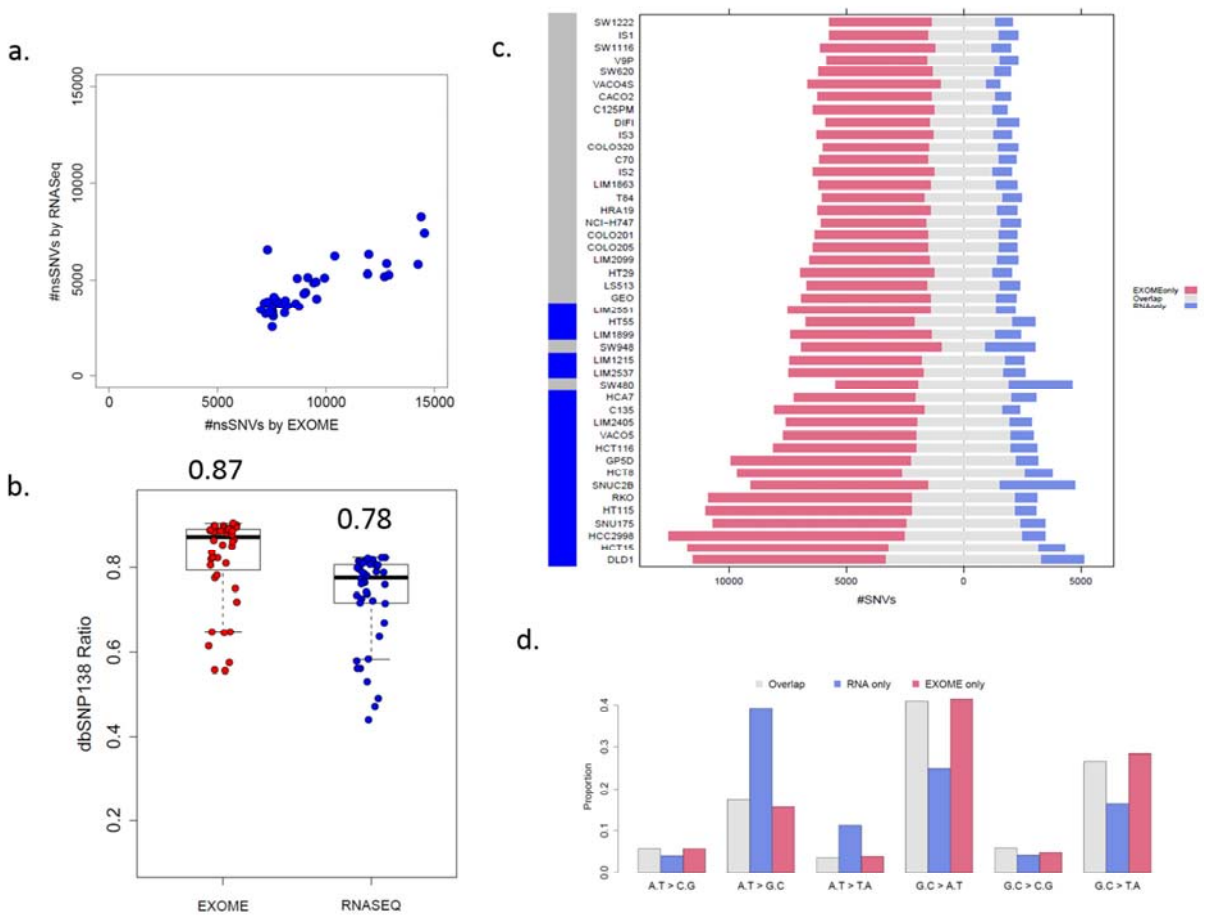
b.

	P-value
All ~ Hyper	*0.0004535*
COSMIC/TCGA ~ Hyper	*0.01765*
dbSNP ~Hyper	*0.07166*
New ~ Hyper	*9.689e-08*



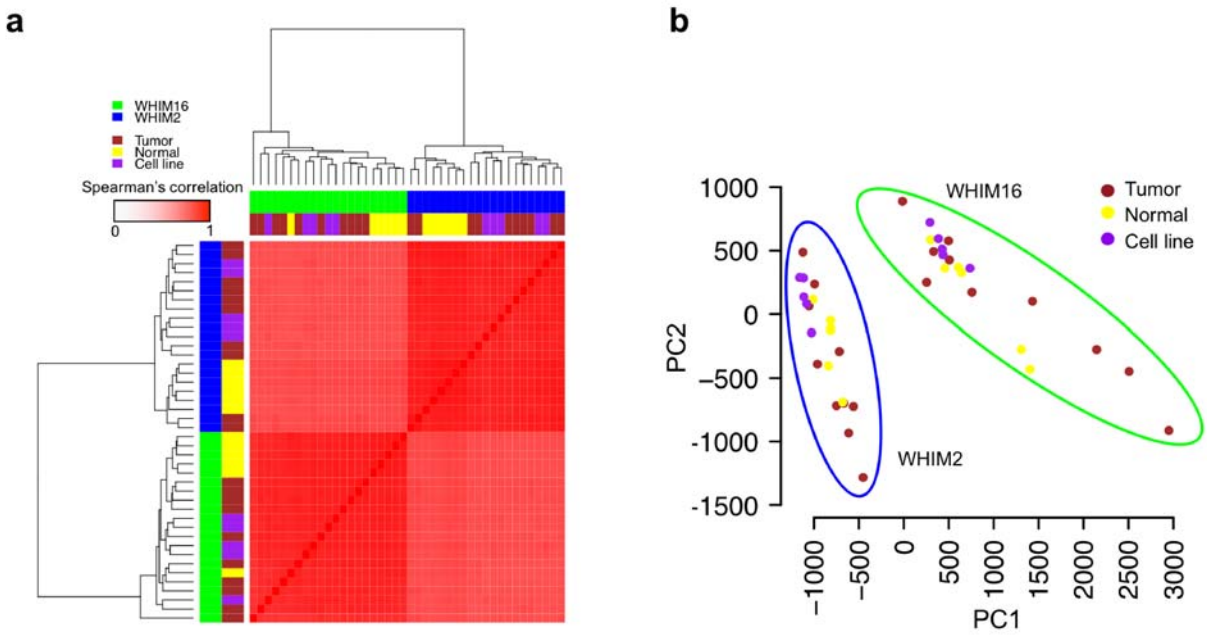
Supplementary Fig 3. Proteomic detection of single nucleotide variants (SNVs) in CRC cell lines. (**a**) Classification of the SNVs detected in individual cell lines based on support from various variant databases. The cell lines are ordered by the number of new variants, then COSMIC/TCGA-supported somatic variants, and then dbSNP-supported variants. Yellow, light blue, and dark red indicate SNVs detected only by exome sequencing based database search, only by RNA-Seq based database search, and by both searches, respectively. Sample hypermutation (Hyper) status is labeled at the bottom (blue, hypermutated; grey, non-hypermutated). (**b**) Association of all detected variants, COSMIC/TCGA supported variants,

dbSNP supported variants and new variants with cell line hypermutation phenotype. *p* for Wilcoxon rank-sum test. (**c**) Distributions of the functional impact scores calculated by MutationAssessor and SIFT for the three categories of SNVs. *p* for t-test.

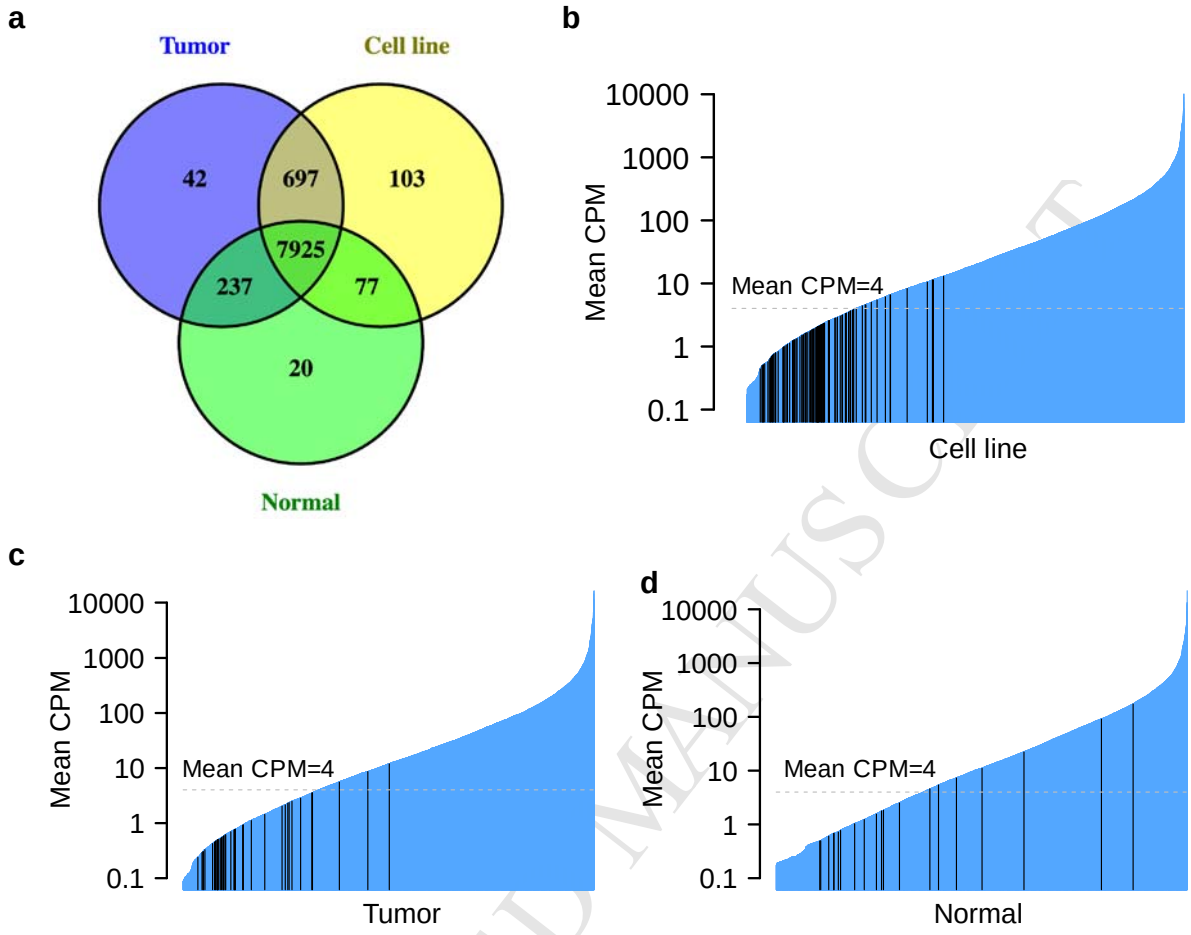


Supplementary Fig. 4. Incorporation of non-synonymous single nucleotide variants (nsSNVs) into a customized database. (**a**) A scatterplot shows the number of nsSNVs detected by exome sequencing (x-axis) and RNASeq (y-axis). Each dot represents a cell line. (**b**) dbSNP rates of ncSNVs detected by exome sequencing (red) and RNASeq (blue) in 44 cell lines. The numbers above the box indicate the median value (**c**) The number of nsSNVs detected by exome sequencing only (red), RNASeq only (blue) and both (grey) for each cell line, ordered by the combined number of unique nsSNVs from the two platforms (from top to bottom). The hypermutation status is labelled on the left (blue, hypermutated; grey, non-hypermutated). (**d**) Mutation spectra for the three nsSNVs categories. Mutational patterns were calculated for

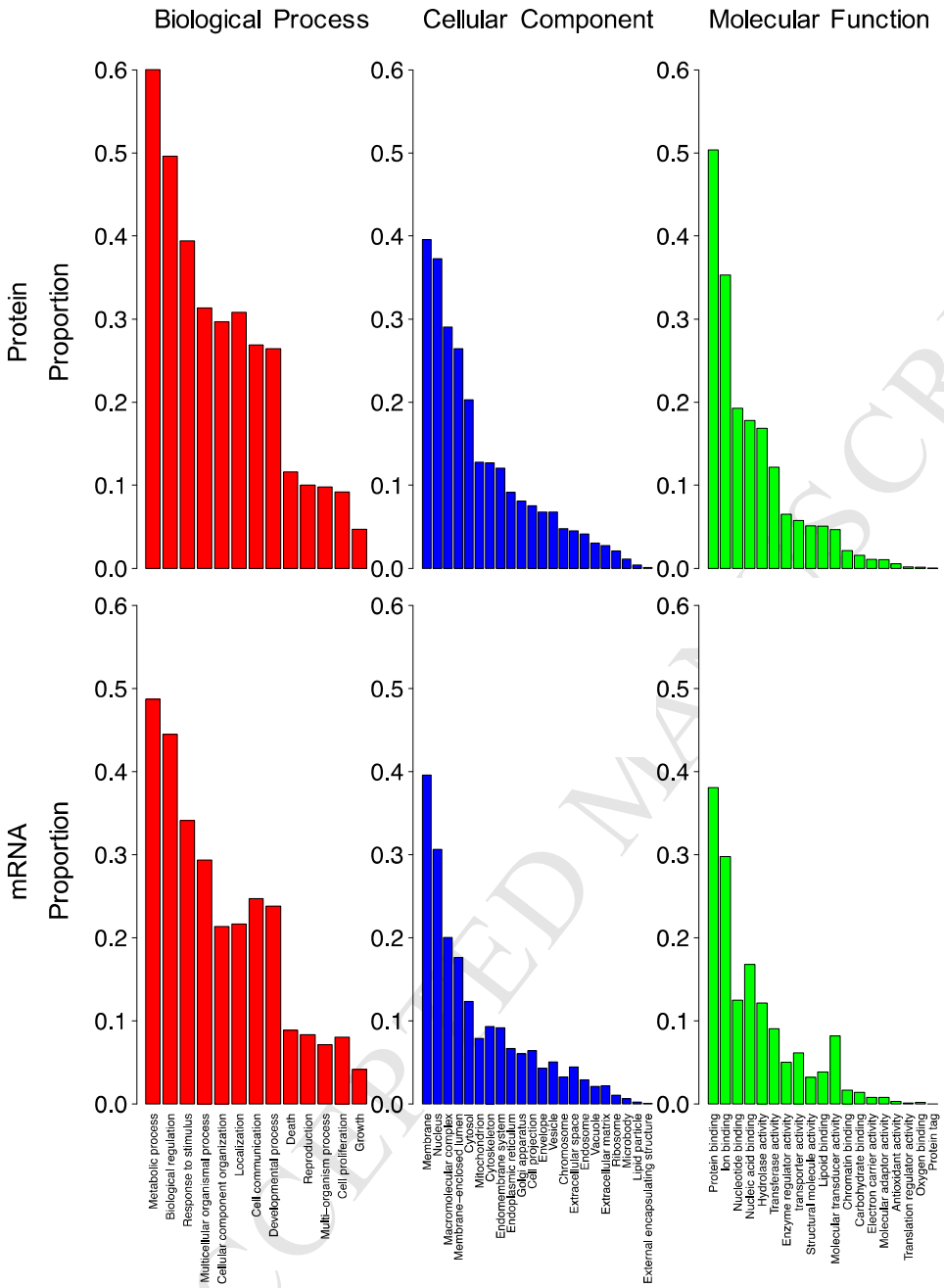
nsSNVs detected by exome sequencing only (red), RNASeq only (blue) and both (grey). Y axis represents the proportion of each nucleotide substitution type. Similar patterns were observed for overlap and exome unique nsSNVs, whereas RNASeq unique nsSNVs showed a different nucleotide substitution patterns. e.g., ~40% of RNA-Seq unique variants show enriched A.T → G.C mutations, which is a hallmark of RNA editing.



Supplementary Fig. 5. Proteomic platform stability. The cell line, tumor, and normal tissue proteomic analysis was performed on the same platform. To evaluate platform stability, we analyzed benchmark quality control (QC) samples (n=42) of basal and luminal human breast tumor xenografts run in alternating order after every five tumor, normal tissue, or cell line samples. (**a**) Heatmap representing the Spearman's correlation between each pair of samples. The green and blue bars represent the WHIM16 and WHIM2 samples, respectively, whereas the brown, yellow, and purple bars represent interstitials within the tumor, normal, and cell line cohorts, respectively. (**b**) Principal component analysis plot. The color scheme is the same as (**a**).

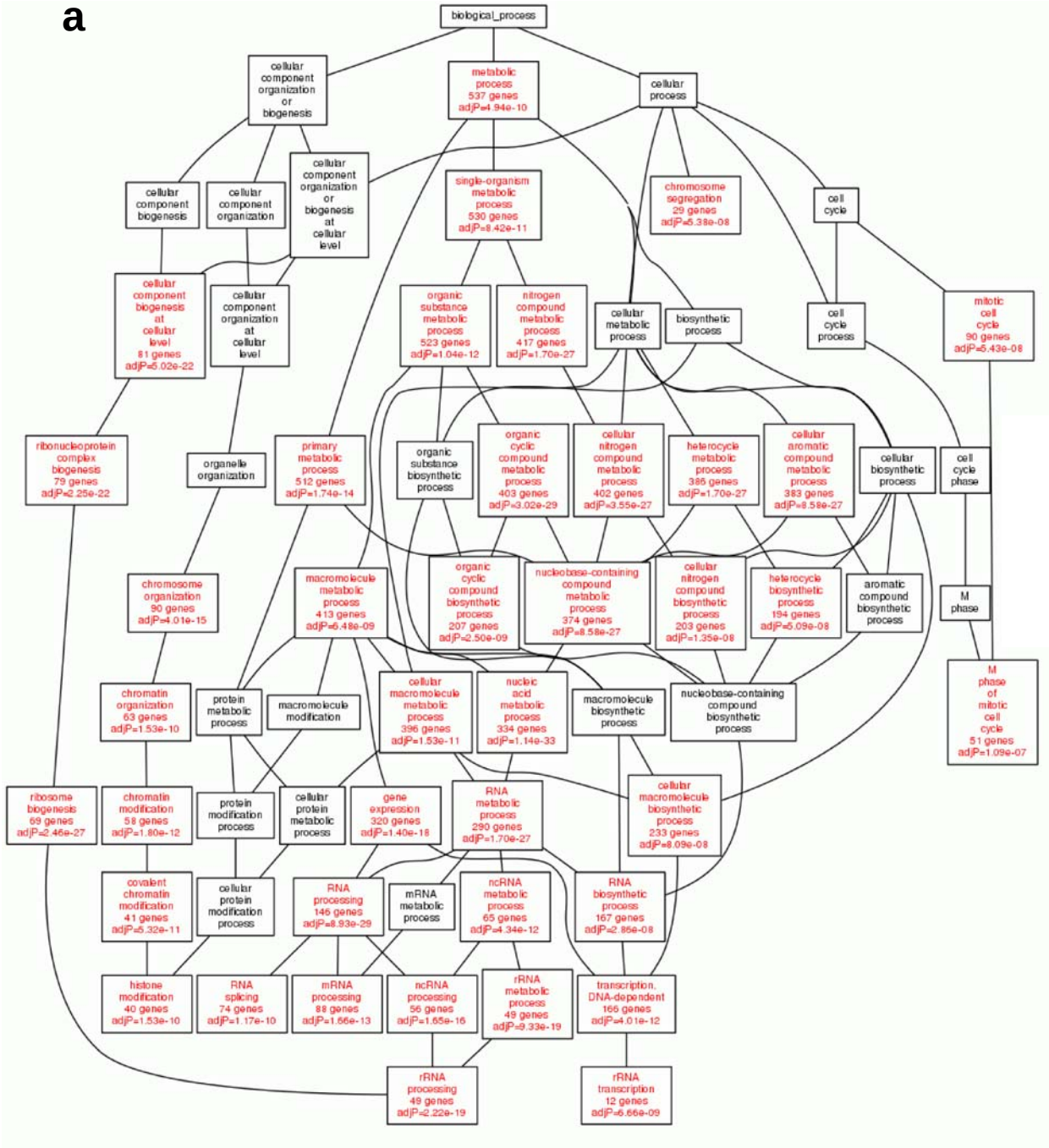


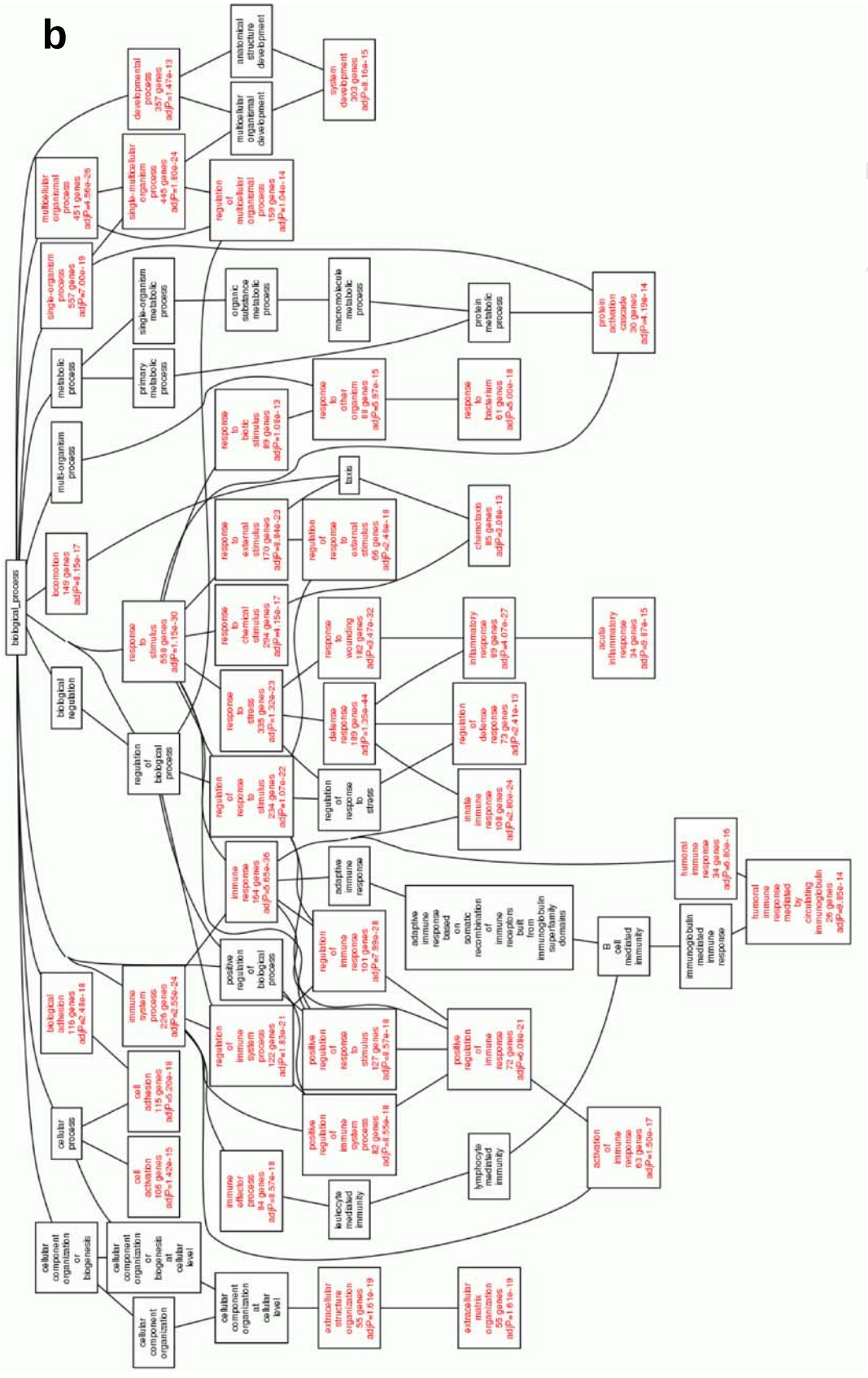
Supplementary Fig. 6. Comparison of protein inventory between cell lines, tumor and normal tissue data. (**a**) Venn diagram comparison for the three data sets. (**b-d**) log-scaled mean CPM distribution for protein inventory of cell line, tumor and normal data. Black lines in the three figures represent 103 cell line-specific proteins, 42 tumor-specific proteins and 20 normal tissue-specific proteins, respectively. 86.4% of cell line-, 92.9% of tumor-, and 65.0% of normal-specific proteins exhibiting <4 spectral counts per million (CPM).



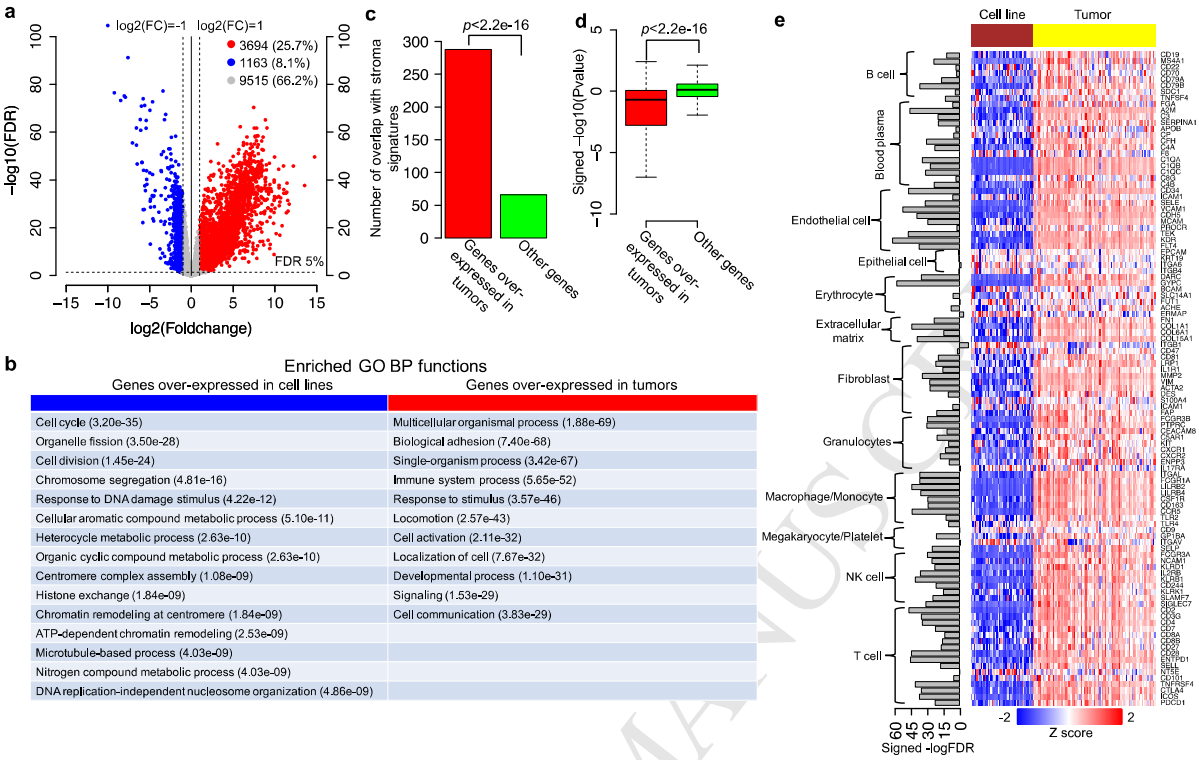
Supplementary Fig. 7. Proteomics (a) and RNA-Seq data (b) according to annotation for major Gene Ontology biological processes, molecular functions, and cellular components.

Terms from GOSlim database.


a
biological_process
cellular component organization or biogenesis
metabolic process 537 genes adjP=4.94e-10
cellular process
cellular component biogenesis
cellular component organization
cellular component organization or biogenesis at cellular level
single-organism metabolic process 530 genes adjP=8.42e-11
chromosome segregation 29 genes adjP=5.38e-08
cell cycle
cellular component biogenesis at cellular level 81 genes adjP=5.02e-22
cellular component organization at cellular level
organic substance metabolic process 523 genes adjP=1.04e-12
nitrogen compound metabolic process 417 genes adjP=1.70e-27
cellular metabolic process
biosynthetic process
cell cycle process
mitotic cell cycle 90 genes adjP=5.43e-08
ribonucleoprotein complex biogenesis 79 genes adjP=2.25e-22
organelle organization
primary metabolic process 512 genes adjP=1.74e-14
organic substance biosynthetic process
organic cyclic compound metabolic process 403 genes adjP=3.02e-29
cellular nitrogen compound metabolic process 402 genes adjP=3.55e-27
heterocycle metabolic process 386 genes adjP=1.70e-27
cellular aromatic compound metabolic process 383 genes adjP=9.58e-27
cellular biosynthetic process
cell cycle phase
chromosome organization 90 genes adjP=4.01e-15
macromolecule metabolic process 413 genes adjP=6.48e-09
organic cyclic compound biosynthetic process 207 genes adjP=2.50e-09
nucleobase-containing compound metabolic process 374 genes adjP=8.58e-27
cellular nitrogen compound biosynthetic process 203 genes adjP=1.35e-08
heterocycle biosynthetic process 194 genes adjP=5.09e-08
aromatic compound biosynthetic process
M phase
chromatin organization 63 genes adjP=1.53e-10
protein metabolic process
macromolecule modification
cellular macromolecule metabolic process 396 genes adjP=1.53e-11
nucleic acid metabolic process 334 genes adjP=1.14e-33
macromolecule biosynthetic process
nucleobase-containing compound biosynthetic process
M phase of mitotic cell cycle 51 genes adjP=1.09e-07
ribosome biogenesis 69 genes adjP=2.45e-27
chromatin modification 58 genes adjP=1.80e-12
protein modification process
cellular protein metabolic process
gene expression 320 genes adjP=1.40e-18
RNA metabolic process 290 genes adjP=1.70e-27
cellular macromolecule biosynthetic process 233 genes adjP=8.09e-08
covalent chromatin modification 41 genes adjP=5.32e-11
cellular protein modification process
RNA processing 146 genes adjP=8.93e-29
mRNA metabolic process
ncRNA metabolic process 65 genes adjP=4.34e-12
RNA biosynthetic process 167 genes adjP=2.85e-08
histone modification 40 genes adjP=1.53e-10
RNA splicing 74 genes adjP=1.17e-10
mRNA processing 88 genes adjP=1.66e-13
ncRNA processing 56 genes adjP=1.65e-16
rRNA metabolic process 49 genes adjP=9.33e-19
transcription, DNA-dependent 166 genes adjP=4.01e-12
rRNA processing 49 genes adjP=2.22e-19
rRNA transcription 12 genes adjP=6.66e-09


b
biological_process
cellular component organization or biogenesis
cellular component organization or biogenesis at cellular level
cellular component organization
cellular component organization at cellular level
extracellular structure organization 55 genes adjP=1.61e-19
extracellular matrix organization 55 genes adjP=1.61e-19
cellular process
cell adhesion 115 genes adjP=5.30e-18
cell activation 106 genes adjP=1.42e-15
immune effector process 84 genes adjP=8.57e-18
leukocyte mediated immunity
lymphocyte mediated immunity
biological adhesion 116 genes adjP=2.48e-18
immune system process 226 genes adjP=2.55e-24
regulation of immune system process 122 genes adjP=1.83e-21
positive regulation of immune system process 82 genes adjP=8.55e-18
positive regulation of response to stimulus 127 genes adjP=8.57e-18
positive regulation of biological process
immune response 164 genes adjP=5.65e-36
regulation of immune response 101 genes adjP=7.99e-28
positive regulation of immune response 72 genes adjP=8.08e-21
activation of immune response 63 genes adjP=1.50e-17
adaptive immune response
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
B cell mediated immunity
immunoglobulin mediated immune response
humoral immune response 34 genes adjP=6.80e-16
humoral immune response mediated by circulating immunoglobulin 26 genes adjP=8.83e-14
regulation of biological process
regulation of response to stimulus 234 genes adjP=1.07e-22
regulation of response to stress
innate immune response 108 genes adjP=2.80e-24
biological regulation
response to stimulus 558 genes adjP=1.15e-30
response to stress 336 genes adjP=1.32e-23
defense response 189 genes adjP=1.35e-44
regulation of defense response 73 genes adjP=2.41e-13
inflammatory response 89 genes adjP=4.07e-27
acute inflammatory response 34 genes adjP=6.87e-15
response to wounding 182 genes adjP=3.47e-32
response to chemical stimulus 294 genes adjP=4.15e-17
response to external stimulus 170 genes adjP=8.84e-23
regulation of response to external stimulus 66 genes adjP=2.48e-18
chemotaxis 85 genes adjP=3.08e-13
taxis
locomotion 149 genes adjP=4.15e-17
multi-organism process
response to biotic stimulus 89 genes adjP=1.08e-13
response to other organism 88 genes adjP=5.97e-15
response to bacterium 61 genes adjP=5.60e-18
metabolic process
primary metabolic process
single-organism metabolic process
organic substance metabolic process
macromolecule metabolic process
protein metabolic process
protein activation cascade 30 genes adjP=4.19e-14
single-organism process 557 genes adjP=7.00e-19
multicellular organismal process 451 genes adjP=4.56e-26
single-multicellular organism process 445 genes adjP=1.80e-24
regulation of multicellular organismal process 159 genes adjP=1.04e-14
developmental process 357 genes adjP=1.47e-13
multicellular organismal development
anatomical structure development
system development 303 genes adjP=8.16e-15

Supplementary Fig. 8. Enriched GO biological process terms for genes overexpressed in cell lines (a) and tumors (b) based on the proteomics data. The boxes with red colored process names, numbers of proteins and adjusted p values represent the enriched GO terms.



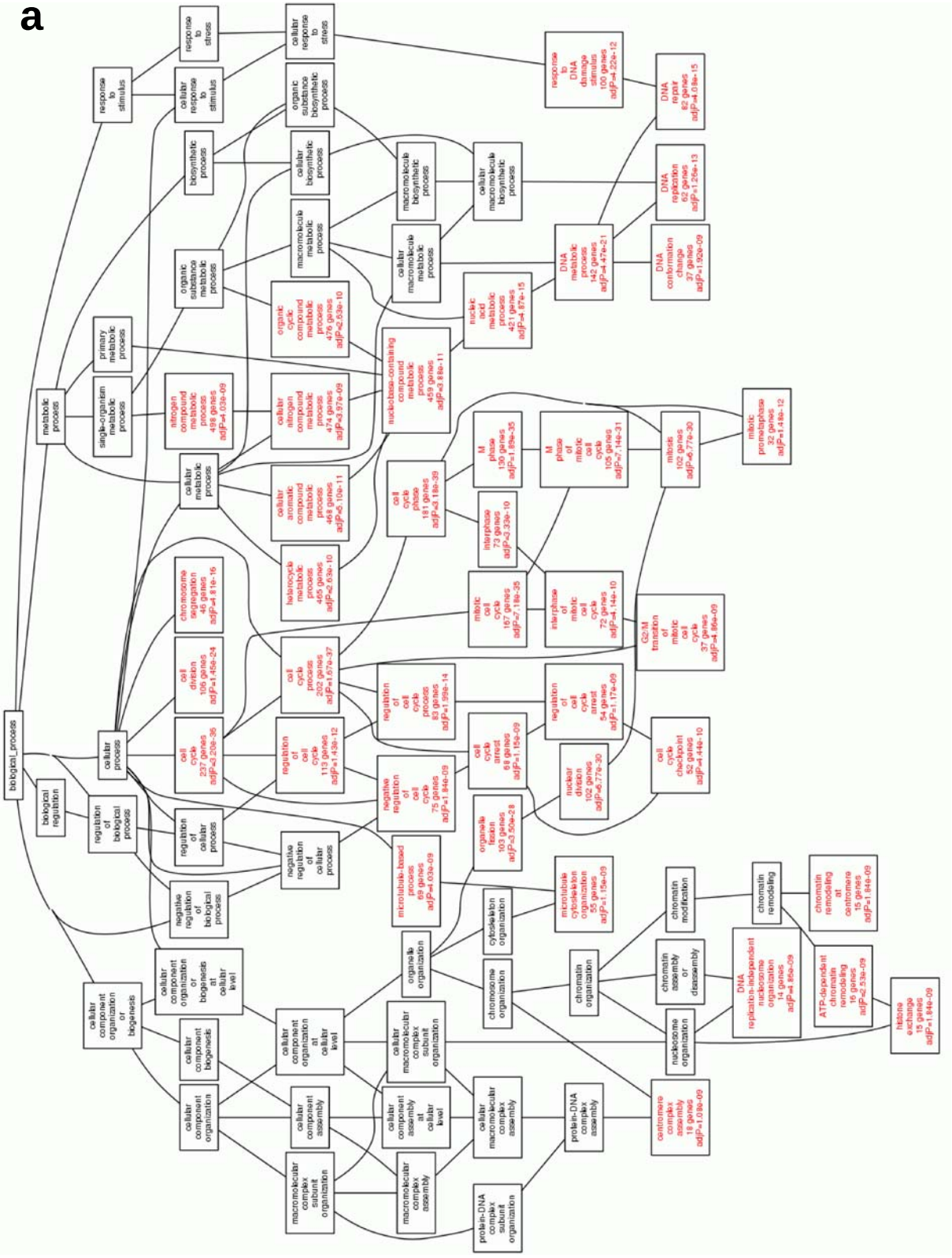
b

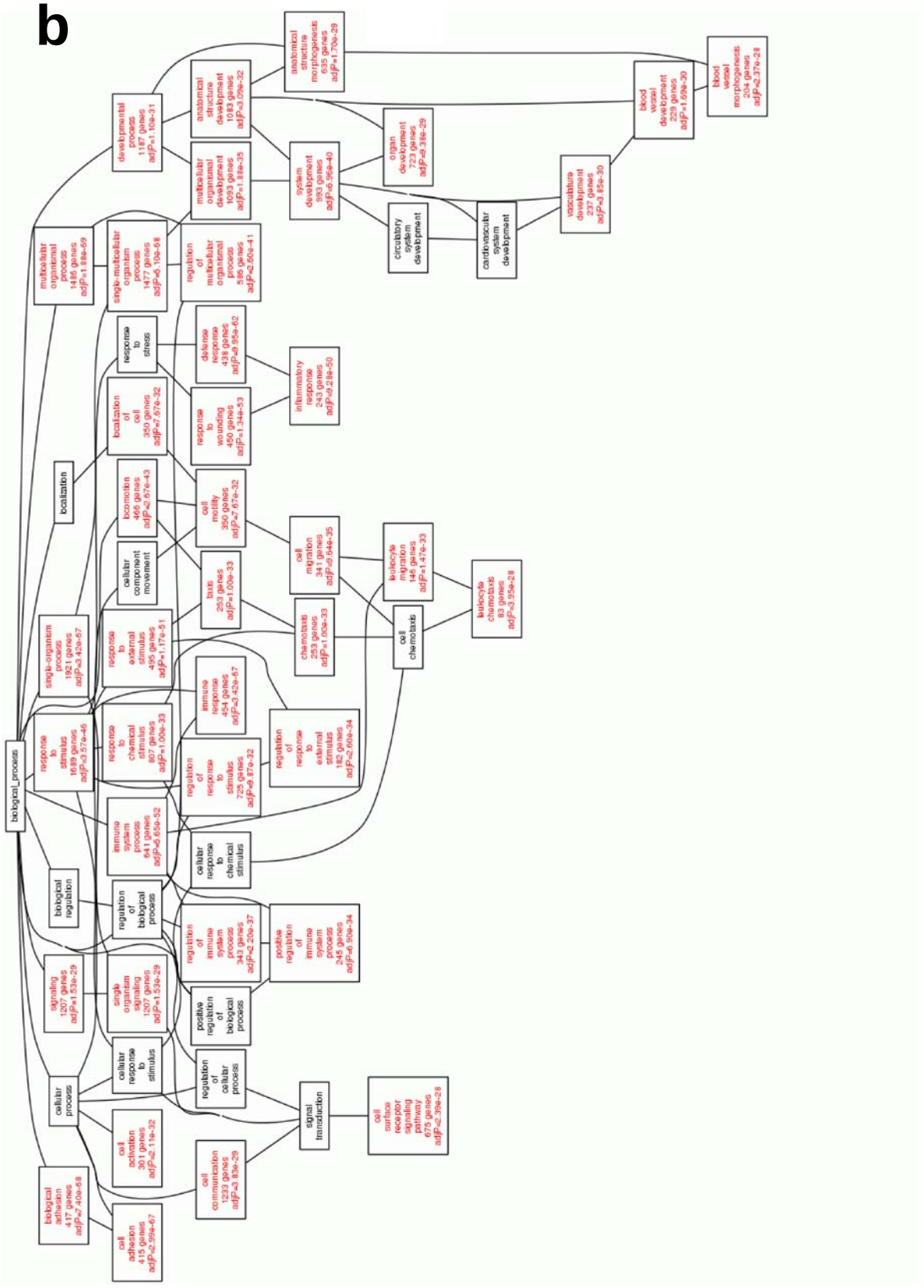
Enriched GO BP functions

Genes over-expressed in cell lines	Genes over-expressed in tumors
Cell cycle (3.20e-35)	Multicellular organismal process (1.88e-69)
Organelle fission (3.50e-28)	Biological adhesion (7.40e-68)
Cell division (1.45e-24)	Single-organism process (3.42e-67)
Chromosome segregation (4.81e-16)	Immune system process (5.65e-52)
Response to DNA damage stimulus (4.22e-12)	Response to stimulus (3.57e-46)
Cellular aromatic compound metabolic process (5.10e-11)	Locomotion (2.57e-43)
Heterocycle metabolic process (2.63e-10)	Cell activation (2.11e-32)
Organic cyclic compound metabolic process (2.63e-10)	Localization of cell (7.67e-32)
Centromere complex assembly (1.08e-09)	Developmental process (1.10e-31)
Histone exchange (1.84e-09)	Signaling (1.53e-29)
Chromatin remodeling at centromere (1.84e-09)	Cell communication (3.83e-29)
ATP-dependent chromatin remodeling (2.53e-09)	
Microtubule-based process (4.03e-09)	
Nitrogen compound metabolic process (4.03e-09)	
DNA replication-independent nucleosome organization (4.86e-09)	

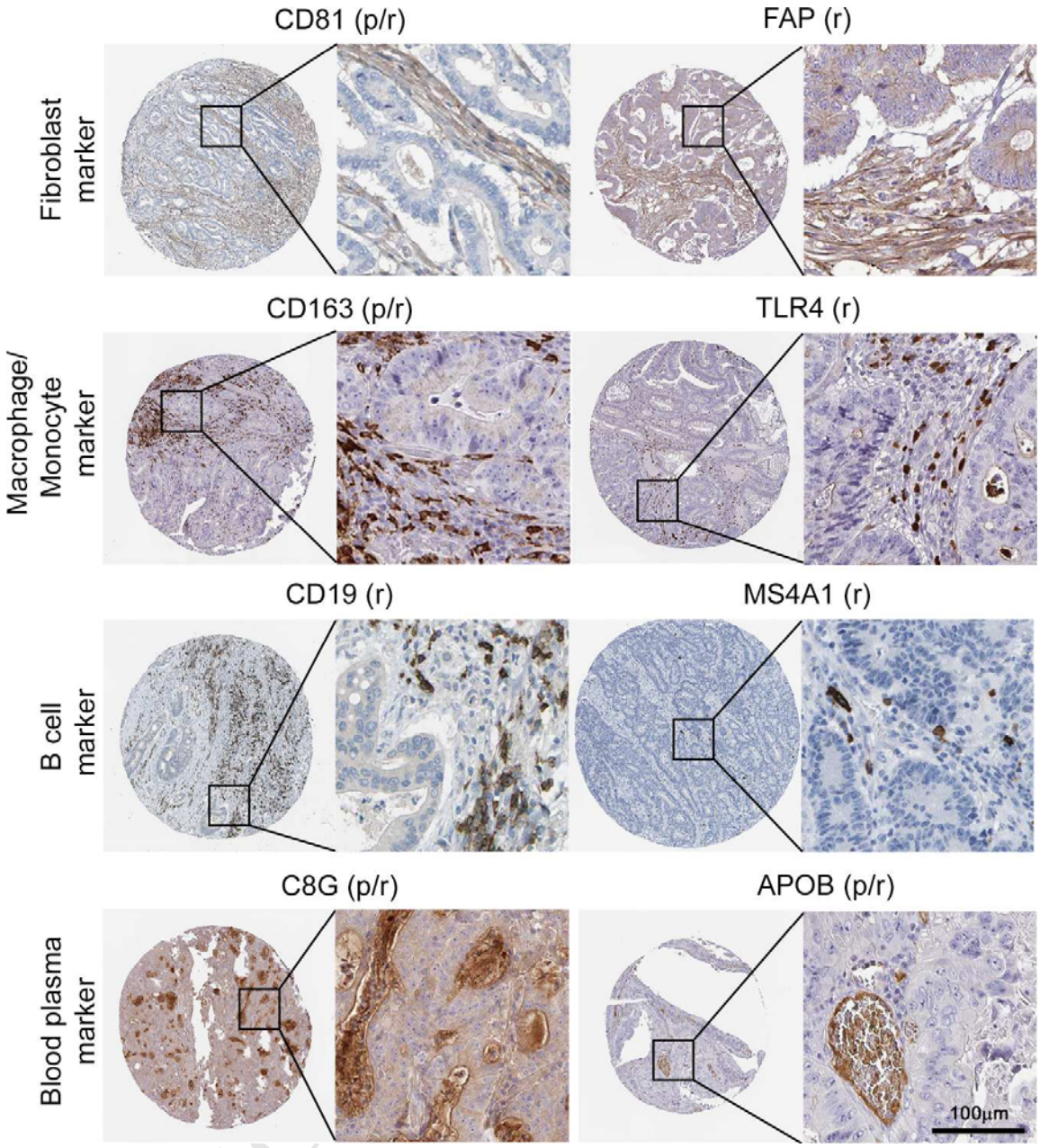
Supplementary Fig. 9. Comparison between mRNA abundance of cell line and tumor data.

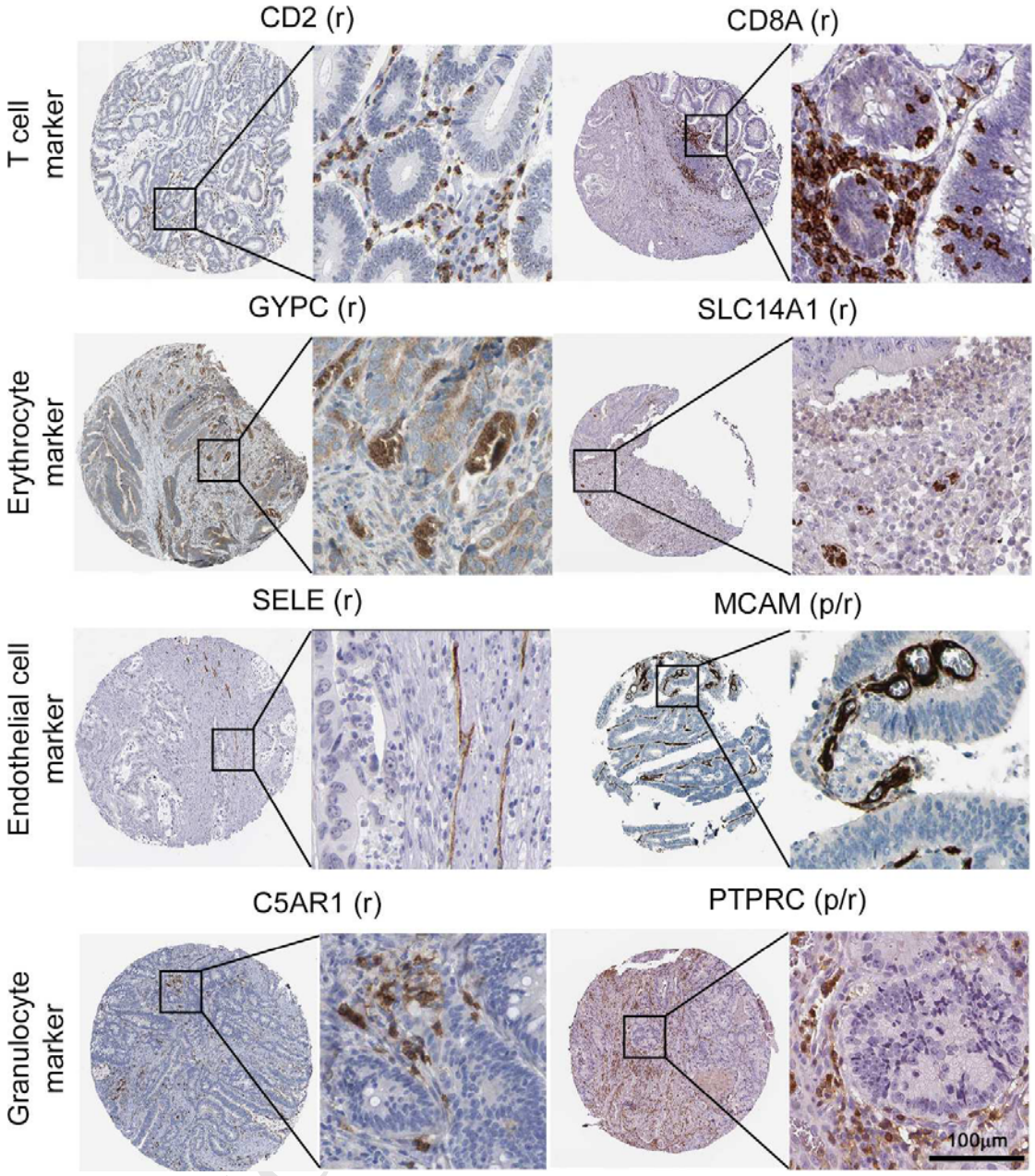
Figure legend is the same as for Fig. 1.

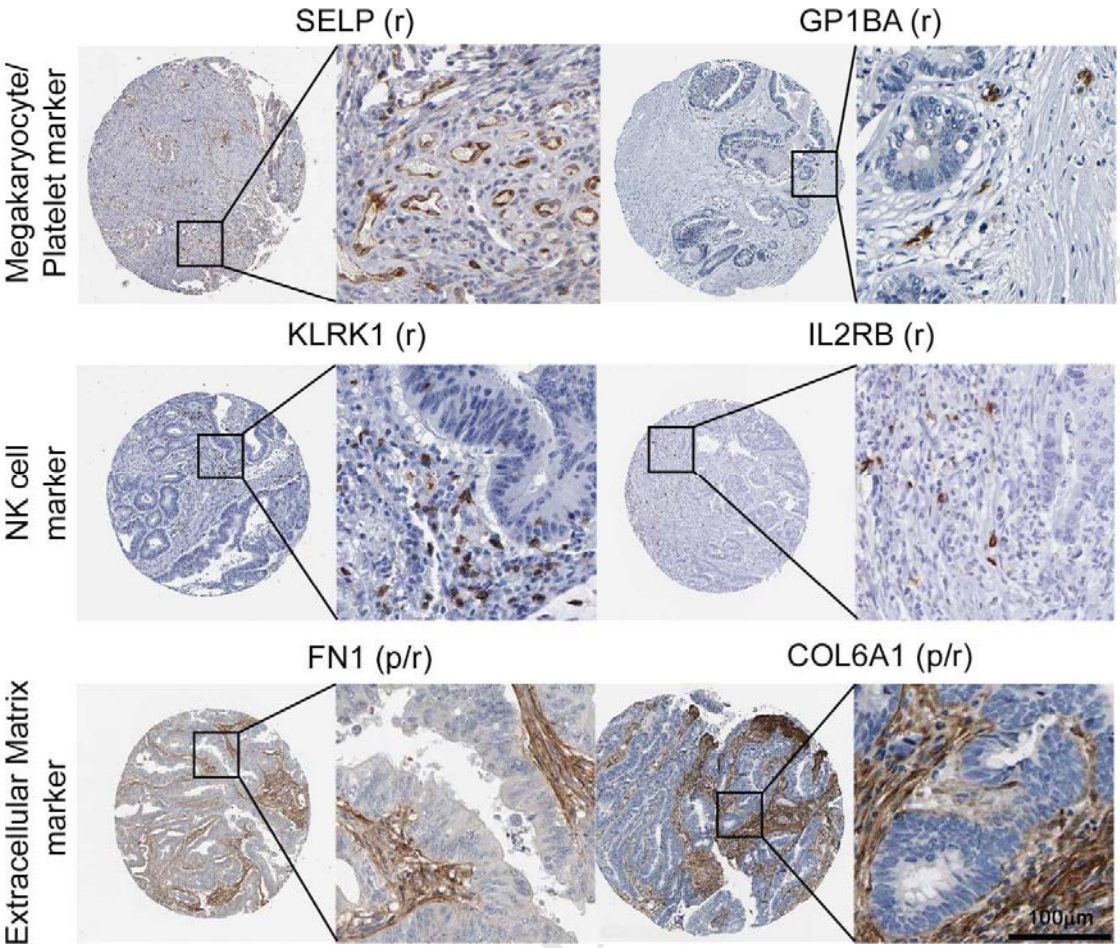

a
biological_process
metabolic process
cellular process
response to stimulus
biological regulation
regulation of biological process
regulation of cellular process
negative regulation of biological process
negative regulation cellular process
cellular component organization or biogenesis
cellular component organization or biogenesis at cellular level
cellular component organization
cellular component biogenesis
cellular component organization at cellular level
cellular component assembly
cellular component assembly at cellular level
macromolecular complex subunit organization
macromolecular complex assembly
protein-DNA complex subunit organization
protein-DNA complex assembly
cellular macromolecular complex assembly
cellular macromolecular complex subunit organization
organelle organization
cytoskeleton organization
chromosome organization
microtubule-based process 69 genes adjP=4.03e-09
microtubule cytoskeleton organization 55 genes adjP=1.15e-09
organelle fission 103 genes adjP=3.50e-28
chromosome segregation 46 genes adjP=4.81e-16
cell division 105 genes adjP=1.45e-24
cell cycle 237 genes adjP=3.20e-36
cell cycle process 202 genes adjP=1.67e-37
regulation of cell cycle 113 genes adjP=1.43e-12
regulation of cell cycle process 83 genes adjP=1.99e-14
negative regulation of cell cycle 75 genes adjP=1.84e-09
cell cycle arrest 58 genes adjP=1.15e-09
nuclear division 102 genes adjP=6.77e-30
regulation of cell cycle arrest 54 genes adjP=1.17e-09
cell cycle checkpoint 52 genes adjP=4.44e-10
Interphase of mitotic cell cycle 72 genes adjP=4.14e-10
mitotic cell cycle 157 genes adjP=7.18e-35
G2/M transition of mitotic cell cycle 37 genes adjP=4.86e-09
interphase 73 genes adjP=2.39e-10
cell cycle phase 181 genes adjP=2.18e-39
M phase 130 genes adjP=1.89e-35
M phase of mitotic cell cycle 105 genes adjP=7.14e-31
mitosis 102 genes adjP=6.77e-30
mitotic prometaphase 32 genes adjP=1.48e-12
nucleobase-containing compound metabolic process 459 genes adjP=3.81e-11
heterocycle metabolic process 466 genes adjP=2.63e-10
cellular aromatic compound metabolic process 468 genes adjP=5.10e-11
organic cyclic compound metabolic process 476 genes adjP=2.63e-10
cellular nitrogen compound metabolic process 474 genes adjP=3.97e-09
nitrogen compound metabolic process 498 genes adjP=4.03e-09
primary metabolic process
single-organism metabolic process
nucleic acid metabolic process 421 genes adjP=4.87e-15
DNA metabolic process 142 genes adjP=4.47e-21
DNA conformation change 37 genes adjP=1.92e-09
DNA replication 62 genes adjP=1.26e-13
DNA repair 82 genes adjP=4.01e-15
response to DNA damage stimulus 100 genes adjP=4.22e-12
cellular macromolecule biosynthetic process
macromolecule biosynthetic process
cellular biosynthetic process
organic substance biosynthetic process
cellular response to stress
response to stress
cellular macromolecule metabolic process
macromolecule metabolic process
organic substance metabolic process
cellular metabolic process
biosynthetic process
cellular response to stimulus
response to stimulus
chromatin organization
chromatin assembly or disassembly
chromatin modification
chromatin remodeling
nucleosome organization
DNA replication-independent nucleosome organization 14 genes adjP=4.85e-09
ATP-dependent chromatin remodeling 16 genes adjP=2.53e-09
chromatin remodeling at centromere 15 genes adjP=1.84e-09
histone exchange 15 genes adjP=1.84e-09
centromere complex assembly 18 genes adjP=1.03e-09


b
biological_process
response to stimulus 1689 genes adjP=3.57e-46
single-organism process 1921 genes adjP=3.42e-67
immune system process 641 genes adjP=6.65e-52
biological regulation
regulation of biological process
signaling 1207 genes adjP=1.53e-29
single organism signaling 1207 genes adjP=1.53e-29
cellular process
cellular response to stimulus
cell activation 301 genes adjP=2.11e-32
cell adhesion 415 genes adjP=2.99e-67
biological adhesion 417 genes adjP=7.40e-68
cell communication 723 genes adjP=3.89e-29
regulation of cellular process
signal transduction
cell surface receptor signaling pathway 675 genes adjP=2.39e-28
positive regulation of biological process
regulation of immune system process 343 genes adjP=2.20e-37
positive regulation of immune system process 245 genes adjP=6.90e-34
regulation of response to stimulus 725 genes adjP=9.87e-32
response to chemical stimulus 807 genes adjP=1.00e-33
response to external stimulus 495 genes adjP=1.17e-51
immune response 494 genes adjP=3.43e-57
regulation of response to external stimulus 182 genes adjP=2.60e-34
cellular component movement
locomotion 466 genes adjP=6.67e-43
localization
localization of cell 350 genes adjP=7.67e-32
cell motility 360 genes adjP=7.67e-32
taxis 253 genes adjP=1.00e-33
chemotaxis 253 genes adjP=1.00e-33
cell chemotaxis
cell migration 341 genes adjP=9.64e-35
leukocyte migration 146 genes adjP=1.47e-33
leukocyte chemotaxis 83 genes adjP=3.95e-28
response to stress
response to wounding 450 genes adjP=1.34e-53
defense response 439 genes adjP=9.85e-62
inflammatory response 243 genes adjP=9.28e-50
multicellular organismal process 1486 genes adjP=1.88e-59
single-multicellular organism process 1477 genes adjP=5.10e-68
regulation of multicellular organismal process 586 genes adjP=2.60e-41
developmental process 1187 genes adjP=1.10e-31
multicellular organismal development 1093 genes adjP=1.88e-35
anatomical structure development 1083 genes adjP=3.09e-32
anatomical structure morphogenesis 635 genes adjP=1.70e-29
system development 993 genes adjP=6.96e-40
organ development 723 genes adjP=9.39e-29
circulatory system development
cardiovascular system development
vasculature development 237 genes adjP=3.85e-30
blood vessel development 229 genes adjP=1.69e-30
blood vessel morphogenesis 204 genes adjP=2.37e-28

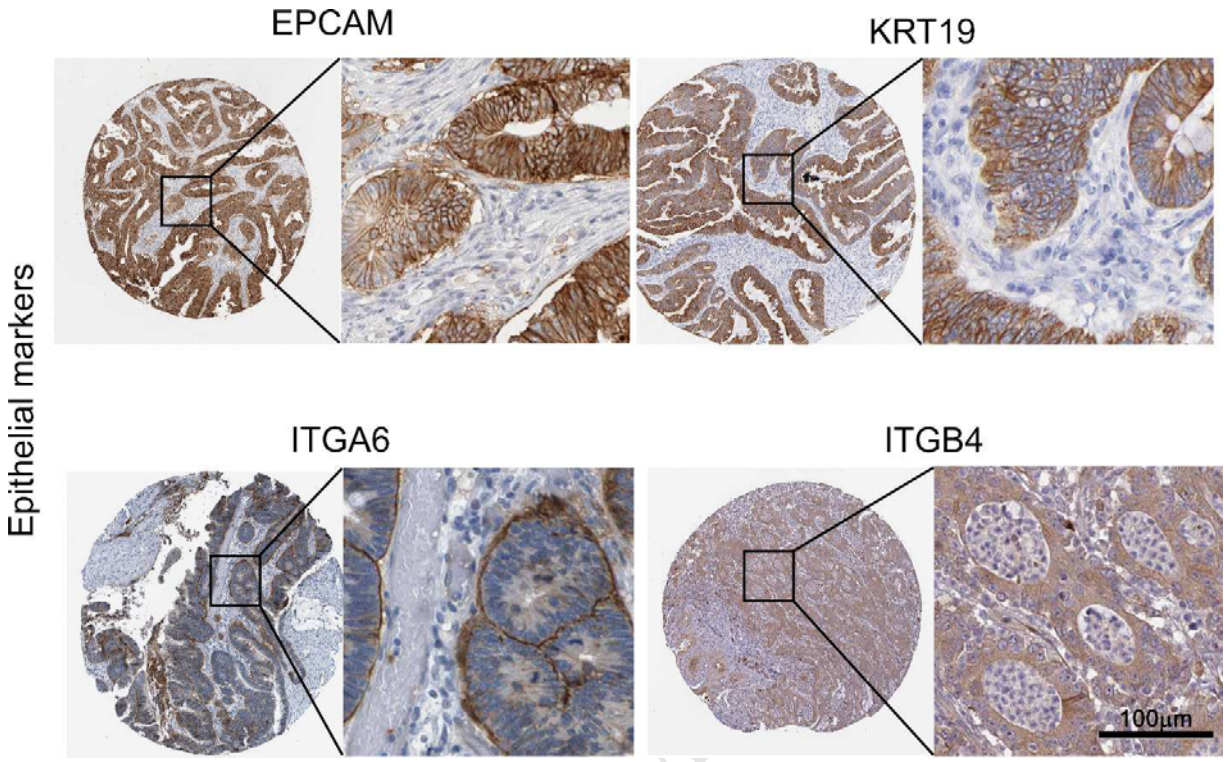
Supplementary Fig. 10. Enriched GO biological process terms for genes overexpressed in cell lines (a) and tumors (b) based on the mRNA data. The boxes with red colored process names, numbers of genes and adjusted p values represent the enriched GO terms.


CD81 (p/r)
FAP (r)
Fibroblast marker
CD163 (p/r)
TLR4 (r)
Macrophage/ Monocyte marker
CD19 (r)
MS4A1 (r)
B cell marker
C8G (p/r)
APOB (p/r)
Blood plasma marker
100µm

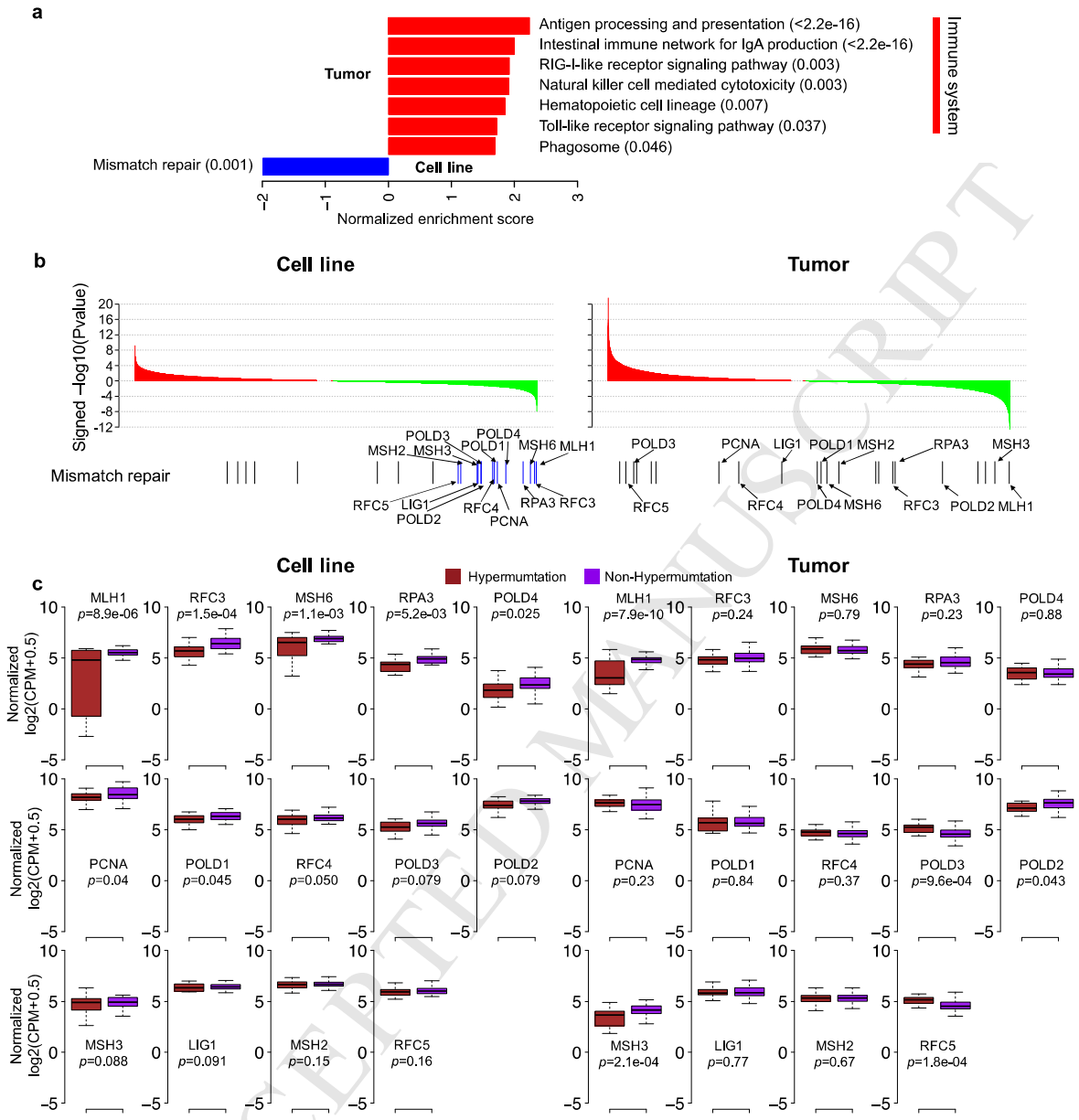

CD2 (r)
CD8A (r)
T cell marker
GYPC (r)
SLC14A1 (r)
Erythrocyte marker
SELE (r)
MCAM (p/r)
Endothelial cell marker
C5AR1 (r)
PTPRC (p/r)
Granulocyte marker
100μm



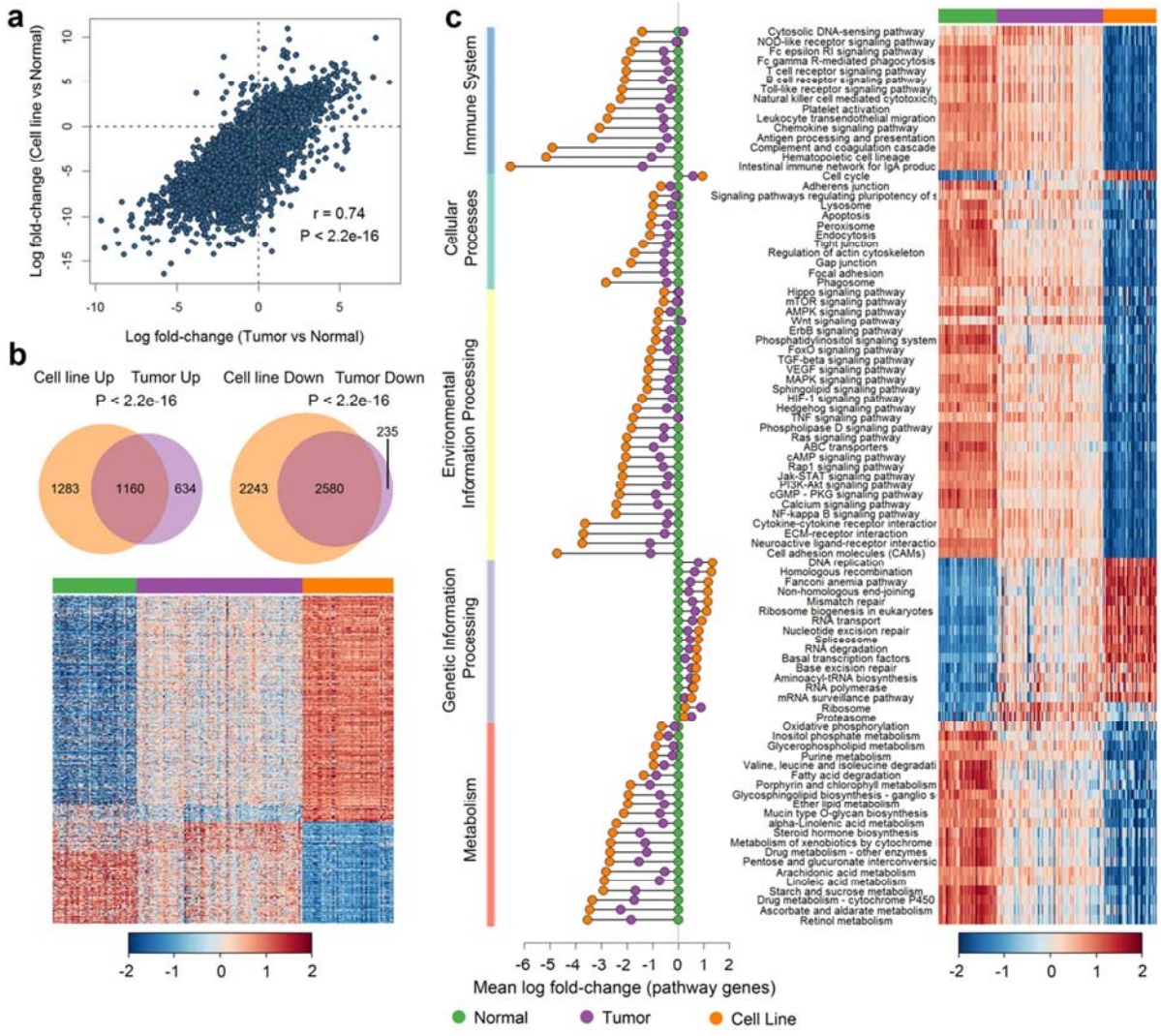
Supplementary Fig. 11. Validation of specificity of blood plasma, extracellular matrix and cell type-specific markers by inspection of the Human Protein Atlas. Exemplar staining images retrieved from the HPA. p/r indicates significance in proteomics and RNASeq data, respectively.



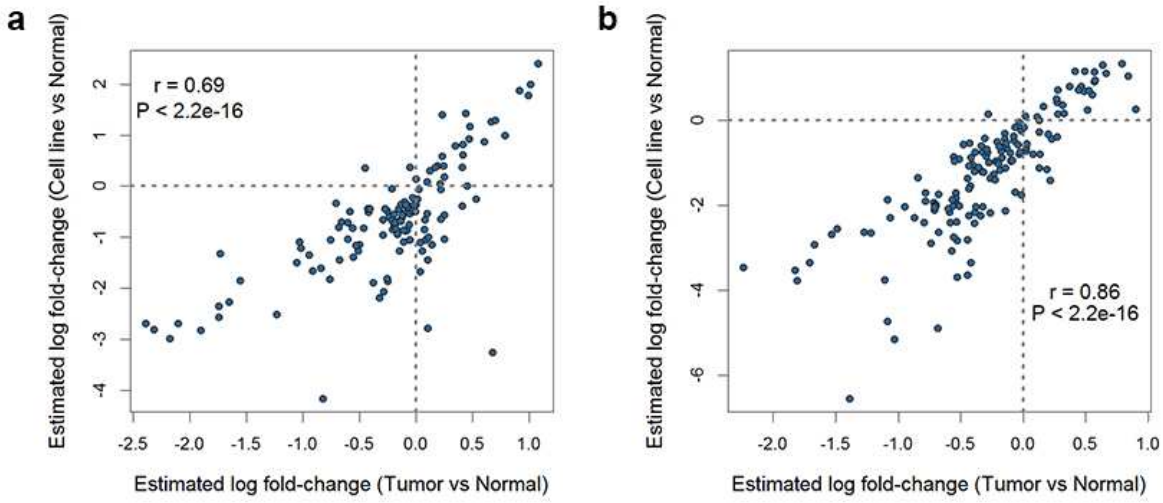
Supplementary Fig. 12. Validation of specificity of epithelial markers by inspection of the Human Protein Atlas. Exemplar staining images retrieved from the HPA.



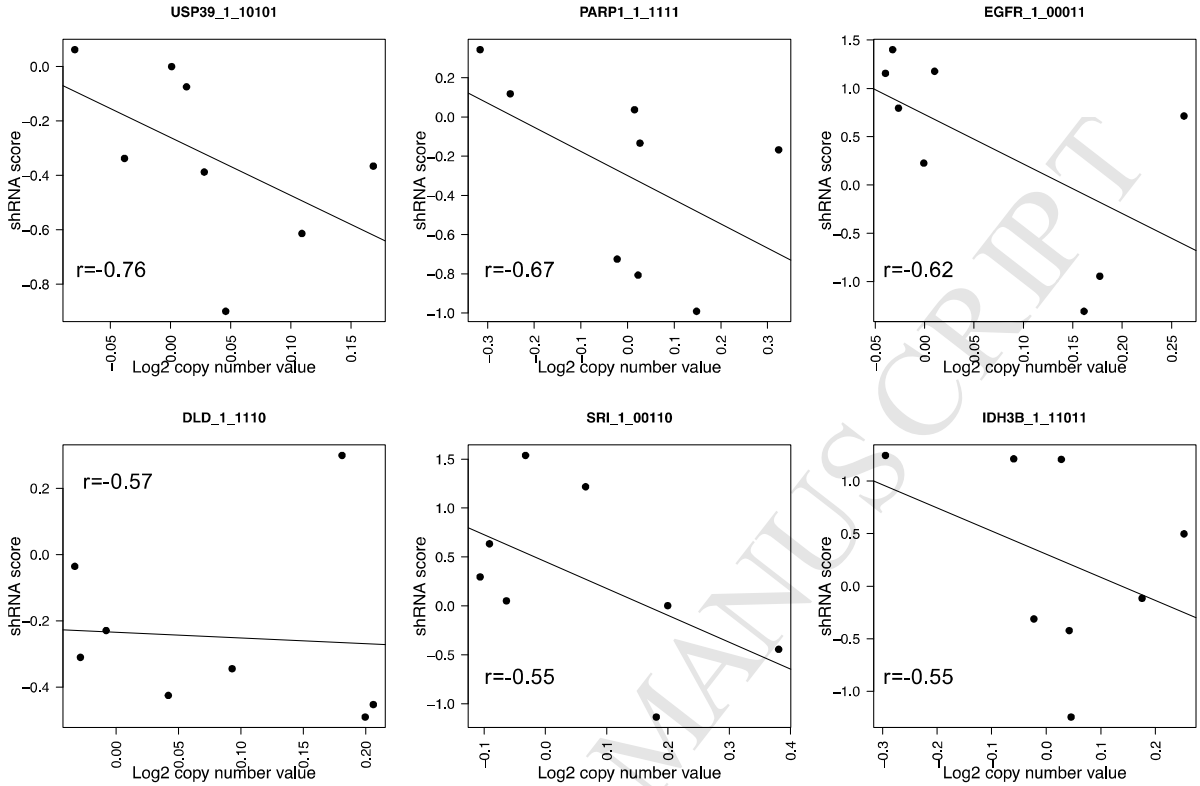
Supplementary Fig. 13. Pathway signatures related to the hypermutated samples for RNASeq cell line and tumor data. Figure legend is the same as for Fig. 2.



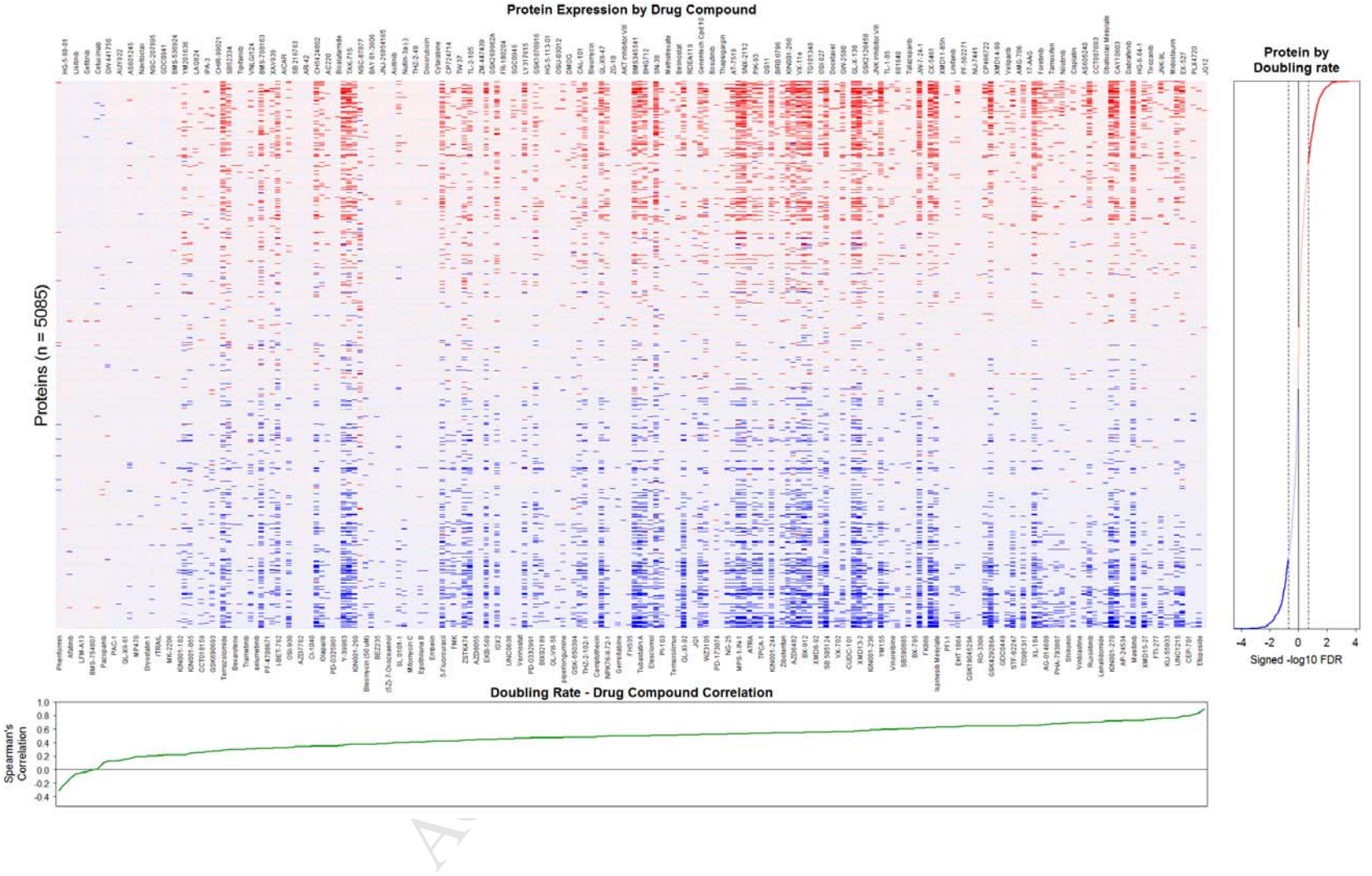
Supplementary Fig. 14. Comparison of cell lines and tumors to normal tissues based on mRNA abundance data. Figure legend is the same as for Fig. 4.



Supplementary Fig. 15. KEGG pathway expression concordance between tumor *versus* normal and cell line *versus* normal for (a) protein and (b) mRNA expression differences.

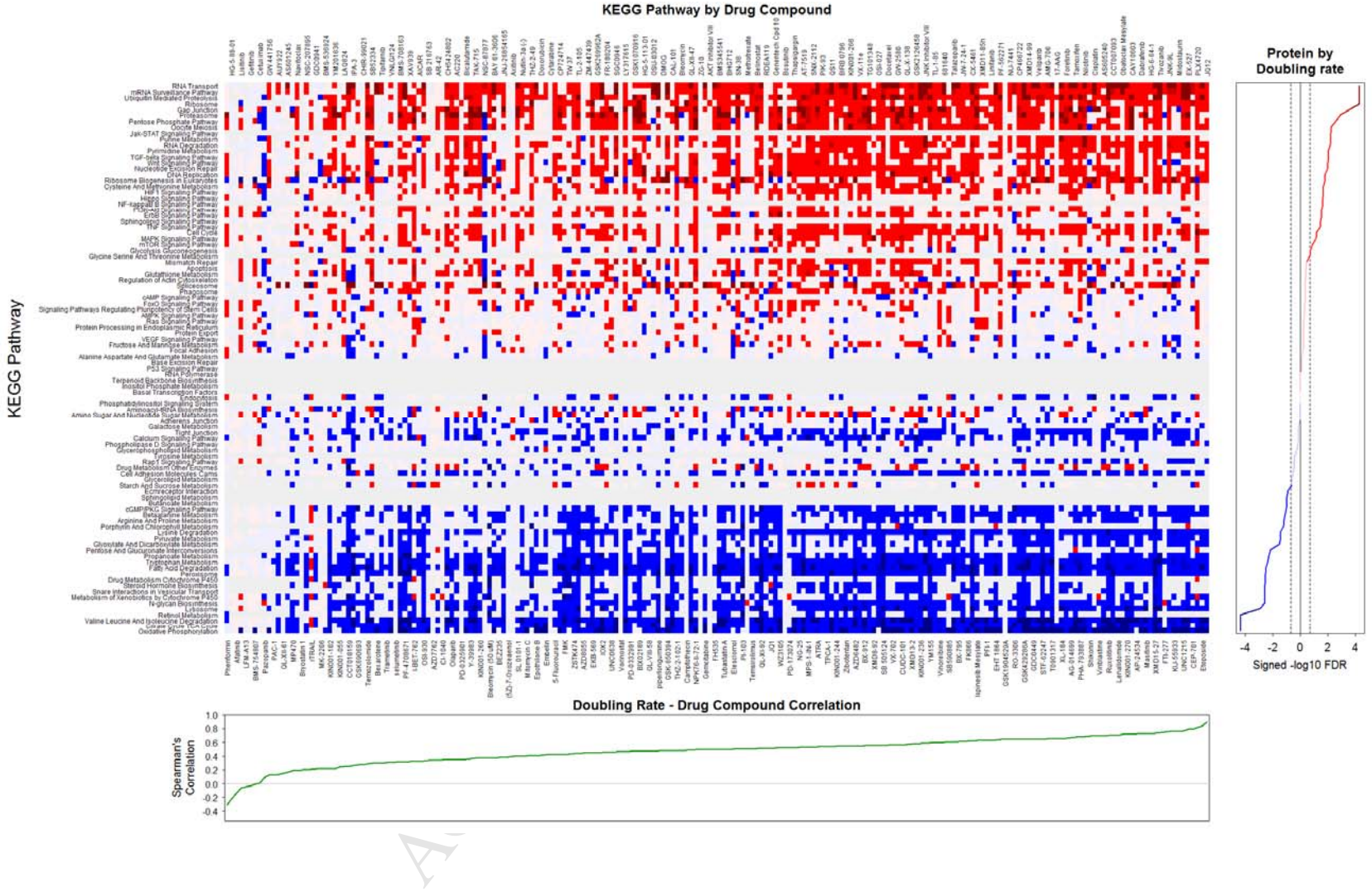


Supplementary Fig. 16. Effect of shRNA for six oncogenes on the proliferation of colon cancer cells based on data from the Achilles study. X-axis represents the log ratio of copy number data and y-axis represents the shRNA score. The correlation was calculated by the Spearman's correlation coefficient.

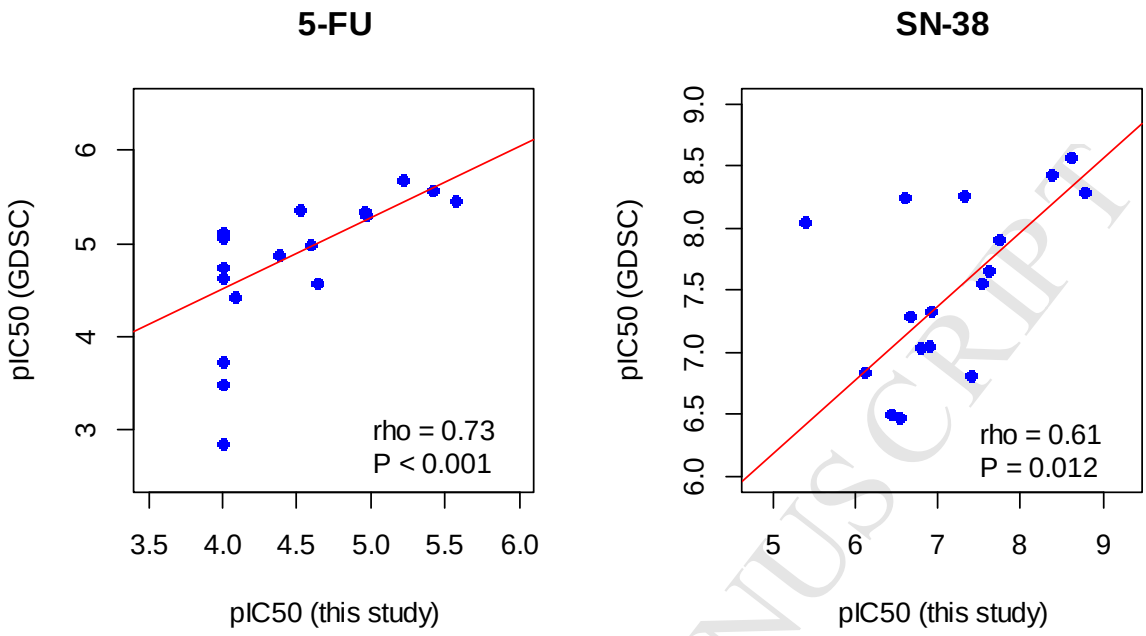


Supplementary Fig. 17. **Heatmap of drug activity-protein associations**. Proteins shown on the vertical axis (Supplementary Table 37). Red and blue coloring, respectively, indicate positive and negative associations between drug pIC_{50} values and protein expression,

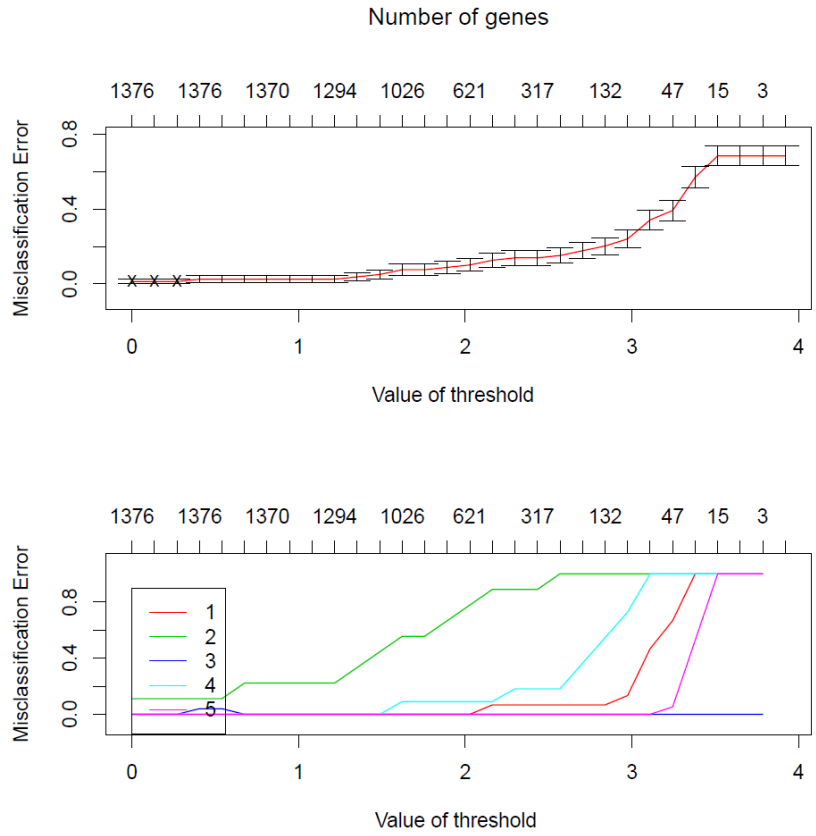
as obtained from voom/limma calculations. The intensity of the coloring corresponds to a range of False Discovery Rate (FDR) values, with the deepest to lightest corresponding to $<10^{-4}$, $<10^{-3}$, < 0.2 and ≥ 0.2. The drugs are ordered by their Spearman correlations for their pIC_{50} values and the cell doubling rate (plotted in the lower panel) and the proteins are ordered by their signed -$\log_{10}$ FDR values for association with cell doubling rate, as obtained from the relevant voom/limma calculation (Supplementary Table 37).



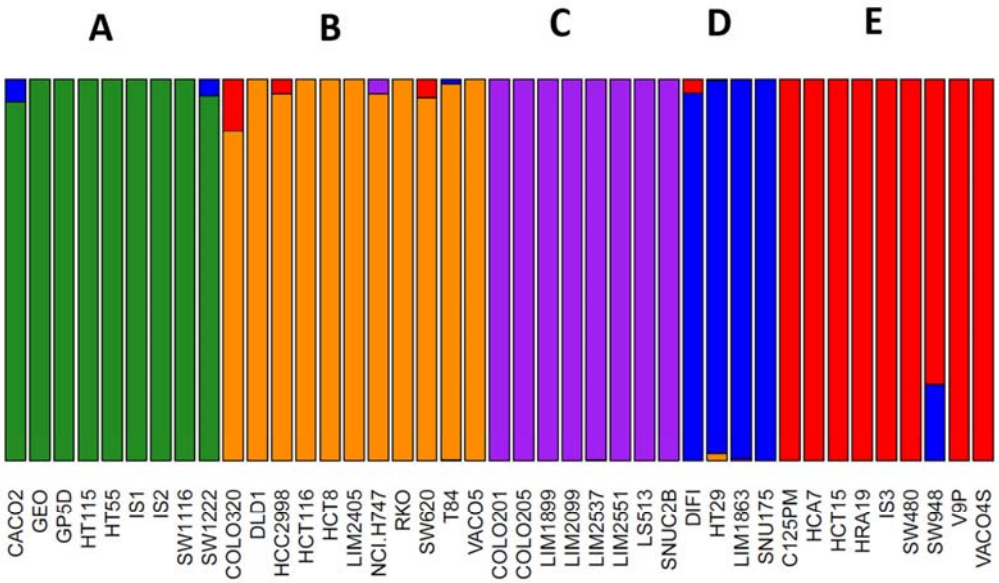
Supplementary Fig. 18. **Heatmap of drug activity-protein KEGG pathway associations.** For each drug, the False Discovery Rates (FDRs) for the KEGG pathways were obtained using the ranked protein list from the relevant voom/limma calculations to assess associations between drug pIC_{50} values and protein expression (Supplementary Table 38). The intensity of the coloring corresponds to a range of FDR values, with the deepest to lightest corresponding to $<10^{-4}$, $<10^{-3}$, < 0.2 and ≥ 0.2. The drugs are ordered by their Spearman correlations for their pIC_{50} values and the cell doubling rate (plotted in the lower panel) and the proteins are ordered by their signed -log10 FDR values for association with cell doubling rate, as obtained from the relevant voom/limma calculation (Supplementary Table 38).



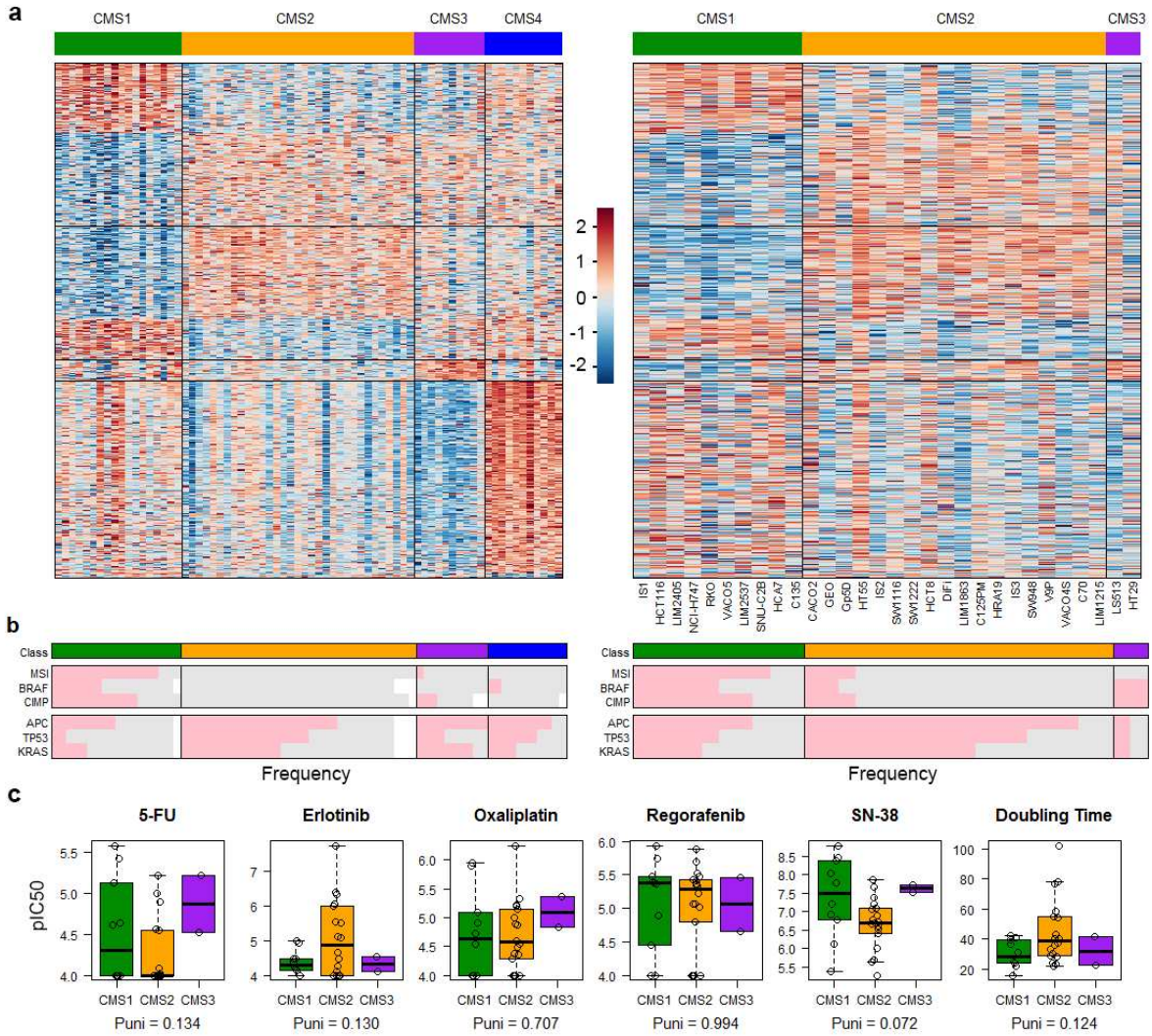
Supplementary Fig. 19. Concordance between GDSC and in-house pIC_{50} data for overlapping drugs.



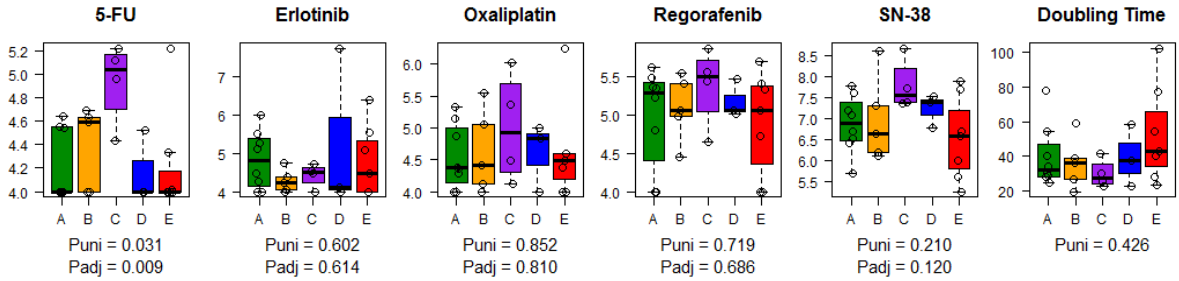
Supplementary Fig. 20. Model for prediction of proteomics subtypes trained on primary tumor assignments. Overall and class-specific misclassification errors are shown for leave-one-out cross-validation. Prediction Analysis of Microarrays (PAM) analysis with increasing values of centroid shrinkage. A set of 1,376 genes provided the minimum cross-validated prediction error.



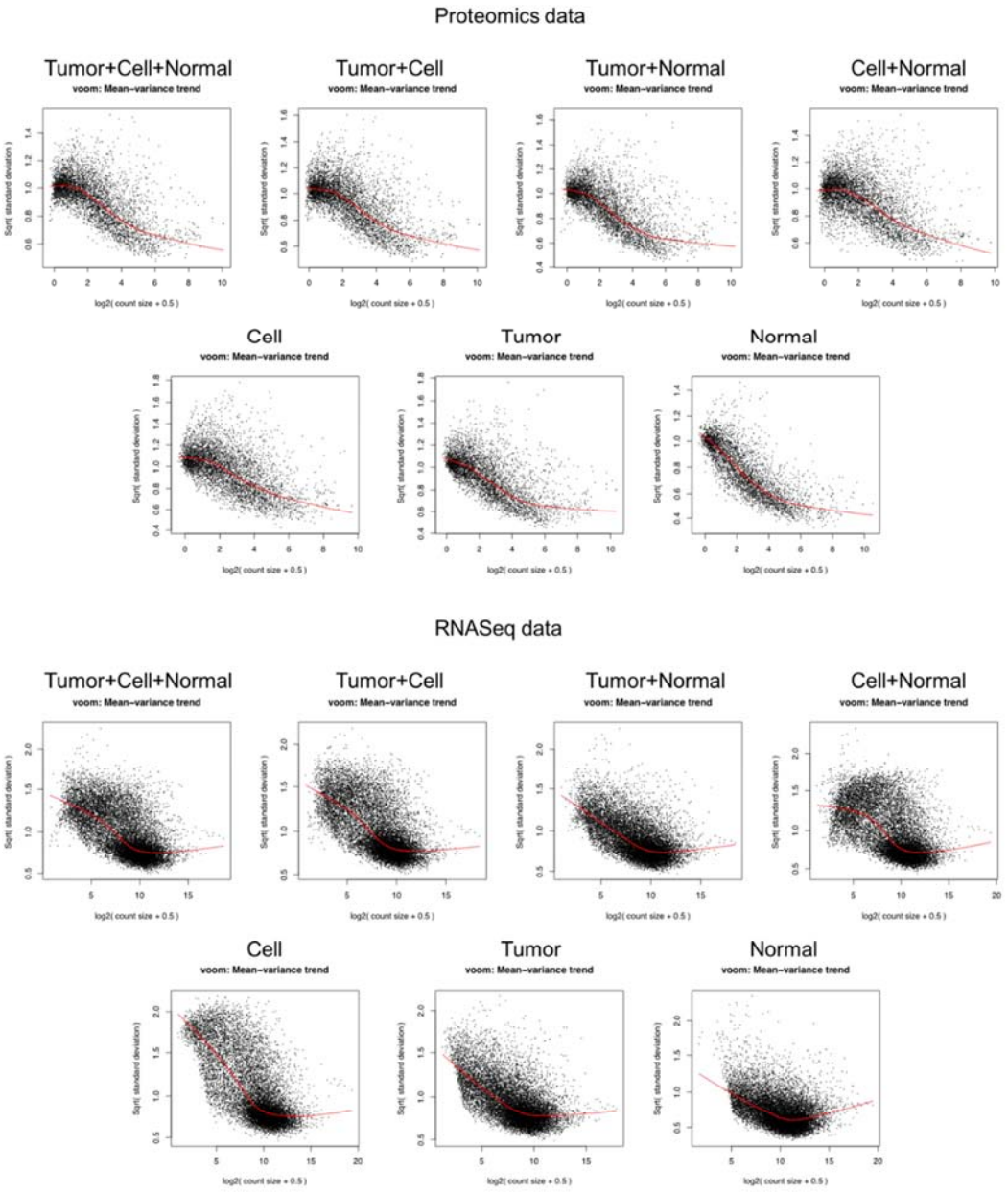
Supplementary Fig. 21. Prediction of proteomics subtypes using 1,376 trained genes. Bar charts indicates the posterior probability of belonging to each proteomics subtype. A. Green; B, Orange; C, Purple; D, Blue; E, Red.



Supplementary Fig. 22. Concordance of mRNA CRC subtypes in cell lines and tumors (a), molecular features (b) and cell line drug response (c). Figure legend is the same as for Fig. 7.



Supplementary Fig. 23. Relationship of proteomics subtypes with drug sensitivity for microsatellite stable cell lines.



Supplementary Fig 24. Mean-variance trend plots based on the different combinations among cell, tumor and normal proteomics data and RNASeq data. The points in the plots represent the quantifiable proteins or mRNAs defined as CPM>20 in 20% of samples for protein spectral counts and CPM>1 in 20% of samples for RNA-Seq counts.